

# 國立交通大學

電信工程研究所

碩士論文

使用取樣點式聲學參數之音素分段

Phonetic Segmentation using Sample-based Acoustic  
Parameters

研究生：林宥余

指導教授：王逸如 博士

中華民國九十九年七月

使用取樣點式聲學參數之音素分段

# Phonetic Segmentation using Sample-based Acoustic Parameters

研究生：林宥余

Student : You-Yu Lin

指導教授：王逸如 博士

Advisor : Dr. Yih-Ru Wang

國立交通大學

電信工程研究所

碩士論文



A Thesis

Submitted to Institute of Communication Engineering

College of Electrical and Computer Engineering

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of

Master

In

Communication Engineering

July 2010

Hsinchu, Taiwan, Republic of China

中華民國九十九年七月

# 使用取樣點式聲學參數之音素分段

研究生：林宥余

指導教授：王逸如 博士

國立交通大學電信工程研究所碩士班



精確的自動語音分段，應用於許多語音辨識系統或是語音合成的研究被認為是有助於提升系統效能的資訊，但是擁有龐大數量的語料庫經由人工準確的標記是相當費時費力，因此本研究以獲得一個精確的音素端點偵測以及自動語音分段系統為目標，以期提升語音辨識或是合成系統的效能。

本論文提出數個取樣點式聲學參數如各頻段信號波封、聲學參數之上升率、頻譜熵以及頻譜 KL 距離，以描述語音信號中各種不同音素之語音特性，加入音素端點偵測以及自動語音分段的系統架構中，再分別針對音素端點以及自動語音分段所選用的基本語音單位訂定目標函數，接著使用前饋式類神經網路多層感知器以半監督式之模型訓練方法來建立音素端點偵測器之模型。最後對於不同語料庫的語句來進行音素端點偵測的實驗與自動語音分段的效能分析。

# Phonetic Segmentation using Sample-based Acoustic Parameters

Student : You-Yu Lin

Advisor : Dr. Yih-Ru Wang

Institute of Communication Engineering  
National Chiao Tung University

## Abstract

Automatic speech segmentation with high precision and accuracy is considered worthwhile in some speech recognition and speech synthesis researches. Manual labeling is the most precise way, but a huge database with manual labeling and segmentation are very time-consuming process. In order to promote the performances of speech recognition/synthesis system, sample-based phone boundary detection and segmentation algorithms are proposed in this paper.

Some sample-based acoustic parameters are first extracted in the proposed method for modeling acoustic features in the spectral of speech signal, including six sub-band signal envelopes, rate of rise, sample-based KL distance and spectral entropy. Then, the sample-based KL distance is used for boundary candidates pre-selection and a target function labeling that specified the state-transitions between different classes which are pre-defined based on the transcription level. Last, a semi-supervised neural network is employed for final phone boundary detection and automatic speech segmentation. Finally, experimental results and analyses for phoneme detection and automatic speech segmentation are discussed with different corpus.

# 致謝

首先，感謝陳信宏老師對學生的照顧，也特別感謝王逸如老師指導我的用心，讓我在研究的過程中學習到真正的做事態度，並且指引著我在研究的過程中不致迷失方向。

在碩士生涯的 2 個年頭，除了老師以外即是由最重要的博班學長們，帶領著我們度過瓶頸，非常感謝根本就是語言學家的性獸博士、雖然很色但是研究方面說話有可靠度的阿德、慢條斯理的合哥、業界闖蕩爽朗的巴金以及常常嚇唬我們的輝哥，謝謝你們在碩班對我的建議與指導!!

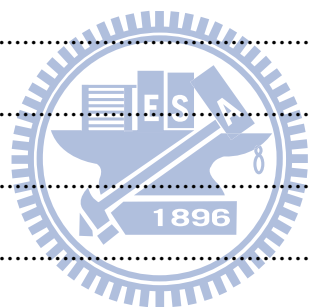
實驗室的生活點滴絕對是刻骨銘心的，上一屆學長美食家普烏、常常哎喲的杜 Q、常常喔喲的小宋、福利社社長小帥哥，同屆夥伴時常共患難的承燁、學妹 MSN 都有的 puma、傻傻的舒舒、工作效率極高的嘴砲小卡、總是在晚班交接的 10、來無影去無蹤的皓翔，學弟少了點勇氣的憨人胖胖、整天想把妹的啟全、講話真的很瞎的小瞎、愛看動漫的豆腐喵、超屌大胖哥、作研究到比我還晚的銘傑、跳舞酷斃了的智障，因為有你們，兩年生活裡的回憶不僅是彩色的，還外加充滿許許多多笑聲、咒罵聲等 3D 立體音效存在我的腦海中，也多謝你們的幫忙讓我的論文能夠順利完成。另外，也感謝系辦江小姐和蘇小姐在工讀時對我的照顧。

接著，感謝我的女朋友總在我心理感覺壓力很大的時刻支持、鼓勵著我，讓我保持信心來解決任何困難。

最後，將此文獻給我的母親，感謝媽媽時時刻刻的懸念，讓在離鄉背井念書的我也能體會到家裡的溫暖。

# 目錄

中文摘要.....	I
Abstract.....	II
致謝.....	III
目錄.....	IV
表目錄.....	VII
圖目錄.....	VIII
第一章 緒論.....	1
1.1 研究動機.....	1
1.2 研究方向.....	2
1.3 相關研究.....	2
1.4 章節概要說明.....	4
第二章 語料庫介紹.....	5
2.1 TIMIT 語料庫簡介.....	5
2.1.1 語音資料.....	5
2.1.2 文字轉寫之人為時間標記.....	6
2.2 國語 TCC-300 語料庫簡介.....	9
2.3 國語 Treebank 語料庫簡介.....	10
2.4 客語語料庫簡介.....	11
第三章 取樣點式之語音聲學參數.....	12
3.1 取樣點式聲學參數之語音特徵.....	12
3.1.1 子頻段信號波封.....	12
3.1.2 上升率.....	15
3.1.3 頻譜熵.....	17



3.1.3 頻譜 KL 距離 .....	18
3.2 使用取樣點式聲學參數之類音素端點自動分段.....	20
第四章音素端點偵測器架構.....	25
4.1 音素端點偵測器架構之設計.....	25
4.1.1 音素端點偵測系統.....	25
4.1.2 自動語音分段系統.....	27
4.2 聲學參數之萃取.....	30
4.2.1 傳統語音聲學參數萃取方式.....	30
4.2.1 取樣點式語音聲學參數萃取方式.....	32
4.3 模型訓練.....	36
4.3.1 多層感知器之類神經網路架構.....	36
4.3.2 反覆疊代.....	42
第五章：實驗結果.....	43
5.1 使用 TIMIT 語料庫之實驗結果.....	43
5.1.1 音素端點偵測實驗結果分析.....	43
5.2 使用國語語料庫之實驗結果.....	52
5.2.1 TCC300 語料庫實驗結果分析.....	52
5.2.2 Treebank 語料庫實驗結果分析.....	55
5.3 使用客語四縣語料庫之實驗結果.....	58
5.3.1 音素端點偵測實驗結果.....	58
5.3.2 自動語音分段實驗結果.....	59
5.4 改良頻譜 KL 距離 .....	60
第六章：結論與未來展望.....	65
6.1 結論.....	65
6.2 未來展望.....	66
參考文獻.....	67

附錄一.....69  
附錄二.....71





# 表目錄

表 2.1：方言之人數分布.....	6
表 2.2：TIMIT 語料庫語句於不同語句類型之分布 .....	6
表 2.3：爆破音對應之短停頓標記符號。.....	7
表 2.4：TCC-300 語料庫檔案統計資料 .....	9
表 3.1：國語語音發音方法的分類表.....	20
表 4.1：類神經網路參數初始設定值.....	41
表 5.1：TIMIT 語料庫的統計資料結果 .....	44
表 5.2：使用音框式計算音素邊界偵測結果的方式的統計結果.....	45
表 5.3：TIMIT 語料庫中發音方法與前後音素不同發音方法之統計資料 .....	47
表 5.4：相鄰音素在相同與不同的發音方法之偵測漏失率.....	48
表 5.5：TIMIT 測試語料中相鄰音素為不同的發音方法之誤報率 .....	51



# 圖目錄

圖 2.1：音素層級之文字轉寫對應於語音信號的人為時間標記.....	7
圖 2.2：國語音節結構圖.....	10
圖 3.1：取樣式語音波封聲學參數範例.....	13
圖 3.2：不同階數之波封檢測器輸出結果.....	14
圖 3.3：取樣式聲學參數之上升率範例.....	16
圖 3.4：取樣式子頻段信號波封聲學參數範例.....	16
圖 3.5：取樣式頻譜熵聲學參數範例.....	17
圖 3.6：取樣式頻譜 KL 距離聲學參數範例.....	18
圖 3.7：不同階數之波封檢測器對頻譜 KL 距離的影響.....	19
圖 3.8：國語語句端點位置自動調整(短停頓)演算法則之範例.....	21
圖 3.9：國語語句端點位置自動調整(摩擦音、塞擦音)演算法則之範例.....	22
圖 3.10：國語語句端點位置自動調整(爆破音)演算法則之範例.....	23
圖 3.11：自動調整國語語句端點位置實驗結果之範例一.....	24
圖 3.12：自動調整國語語句端點位置實驗結果之範例二.....	24
圖 4.1：使用多層感知器架構之音素端點偵測器.....	26
圖 4.2：使用多層感知器架構之自動語音分段系統流程圖.....	27
圖 4.3：音節層級目標函數之轉移狀態圖.....	28
圖 4.4：聲/韻母層級目標函數之轉移狀態圖.....	29
圖 4.5：類音素層級目標函數之轉移狀態圖.....	29
圖 4.6：調整音素候選端點之範例.....	33
圖 4.7：利用候選端點將語音信號分割成片段的示意圖.....	34
圖 4.8：聲學參數抽取演算法的系統架構圖.....	35
圖 4.9：神經元輸入輸出關係圖.....	37

圖 4.10：雙曲正切函數之激發函數曲線圖.....	37
圖 4.11：多層前饋式類神經網路結構範例.....	38
圖 4.12：音素端點偵測器模型反覆疊代之流程圖.....	42
圖 5.1：音素端點偵測器於 TIMIT 語料庫誤報率與偵測漏失率之對應曲線圖 .....	45
圖 5.2：音素端點偵測器實驗結果與人為標記之絕對偏差值直方圖.....	46
圖 5.3：音素端點偵測前後音素為摩擦音之範例.....	48
圖 5.4：音素端點偵測前後音素為鼻音之範例.....	49
圖 5.5：音素端點偵測前後音素為母音之範例.....	49
圖 5.6：音素端點偵測前後音素為靜音之範例.....	50
圖 5.7：音素端點偵測誤報率分析之範例.....	51
圖 5.8：國語語句自動語音分段之範例一.....	52
圖 5.9：國語語句自動語音分段之範例二.....	53
圖 5.10：實驗方法與人為標記位置之誤差在不同絕對偏差值的包含比率直方圖.....	54
圖 5.11：實驗方法與人為標記位置之誤差以發音方法對應不同絕對偏差值的包含比率直方圖 .....	55
圖 5.12：實驗方法與人為標記位置之誤差在不同絕對偏差值的包含比率直方圖.....	56
圖 5.13：不同音節結構實驗結果與人為標記位置之誤差在不同絕對偏差值的包含比率直方圖 .....	57
圖 5.14：偵測客語語句音素端點之範例.....	58
圖 5.15：客語語句自動語音分段之範例一.....	59
圖 5.16：客語語句自動語音分段之範例二.....	59
圖 5.17：疊代至收斂後，對應於每個頻帶的加權值.....	63
圖 5.18：加權值為根據不同參數偏權值與調整挑選候選端點臨限值的結果.....	64
圖 5.19：加入加權頻譜 KL 距離於 TIMIT 測試語料誤報率與偵測漏失率之對應曲線圖 .....	64

# 第一章 緒論

## 1.1 研究動機

現今語音技術的發展與語料庫之間其關係密不可分。無論是用於語音辨識或是語音合成的方面，一個具備完整性且高使用價值性的語料庫是非常重要的。然而，對於語音辨識以及語音合成系統，完整性、高使用價值性是依照語料庫內所含有的資訊來評量，其語音檔案的聲音品質、語句內文字檔案的設計規劃以及最重要的語音信號與文字之時間標記等等，這些資訊皆為評量語料庫價值的重要指標。而音素是語音當中最小的單位，且每種語言中某些部分音素的特性是非常相似的，音素之間也能透過適當結合進而組成音節、詞甚至是片語。

正確音素分段位置在語音辨認的研究中可以提升辨識模型的可靠度與統計上一致性進而提升辨識率[1]，也扮演著語音合成方面合成聲音品質提升的重要因素之一。在全球有人工時間標記音素位置的語料庫不多，最著名的是 TIMIT 語料庫，其同時也是本論文中所使用的語料庫，但是一個大型的連續語音資料庫，使用人工標記音素位置的方式，不僅非常耗時且人工時間標記音素位置也伴隨著一個缺點，就是以人工做標記的動作時，會因為主觀上認定音素端點位置不同，使得標記的位置缺乏一致性，因此一個能夠自動標記且具有精確音素分段的語料庫是非常重要的。

在語音信號處理中，自動音素之分段是一個非常重要的問題，儘管在過去有非常多自動音素分段的研究[2]，一個具有高精準度的自動音素分段演算法，仍是一個可待持續研究的課題。故在本研究中提出取樣點式 (sample-based) 音素端點偵測方法的架構，來提高音素端點偵測及自動分段位置的精準度與準確度之效能。

## 1.2 研究方向

在本篇論文中，將以獲得一個良好的音素端點偵測以及自動語音分段系統為目標，因此本研究結合語言學家所提出的 (Articulation Parameter, AP)，並提出取樣點式音素端點偵測方法的架構，利用數個頻段來區分不同發音特徵之方法，應用於將語音信號做分段可提高時間解析度由音框進一步地精準至取樣點，並在此提出一些取樣點式的聲學參數以用於描述不同語音信號變化時的聲學特性，依此來調整音素位置之標記。

接著本研究利用類神經網路的多層感知器結構 (Multi-Layer Perceptron, MLP) 其自我調適的能力、非線性的運算、具有學習能力等特性，來建立音素端點偵測的模型。我們提出之取樣點式音素端點偵測方法的架構，將語音信號萃取出取樣點式的聲學參數，對語音來進行音素端點的偵測，並利用端點偵測後的結果來觀察其語音信號的變化及自動音素端點分段結果的分析。另外，語音信號的發音特徵應是可以用於所有語言的，意謂著可利用音素端點偵測器來對不同語言之語句進行音素的端點偵測。因此最後本研究以取樣點式音素端點偵測方法的架構應用至國語及客語語料庫，並進行實驗跨語言的音素端點偵測之情況。

## 1.3 相關研究

在過去一些自動音素分段與偵測的研究中，主要可分為以數學模型為基礎 (Model-based) 及以量測為基礎 (Metric-based) 或是上述兩種方法結合。

在 Model-based 方法中，最常被使用的就是以概似法則訓練的隱藏式馬可夫模型 (Maximum Likelihood-trained Hidden Markov Model, ML-trained HMM) 做自動語音分段，其效能可在正負 20 ms 之內佔有 90% 的包含比率 (inclusion rate)，而傳統 HMM 是以整段語句所得到最大相似度函數為訓練準則，故其自動分段之位置並非為最佳之音節或音素端點。近年來有學者提出一些方法，其中以最小邊界錯誤 (Minimum Boundary Error, MBE) 為訓練準則之 HMM[3]，就使用自動與給定之已知端點間誤差最小化作為 HMM 模型之訓練準則，在 TIMIT 語音語料庫中，MBE-HMM 自動分段之邊界與人工標記音素端點誤差範圍 10

ms 之內的比率高達 79.75%，與傳統 ML-trained HMM 模型其百分比 71.23% 相比，提昇許多；然而其自動音素分段位置只有 7.89% 的邊界在人工標記位置誤差 20 ms 之外。此外，也可進一步使用其它圖形識別的方法如支撐向量機[4] (Support Vector Machine, SVM)、類神經網路[5] (Neural Network, NN)，皆可用來對 HMM 之自動分段位置再作進一步地修正以獲得更好的結果。

而在 Metric-based 方法中，我們知道語音信號在一個音素中穩定的信號，其聲學參數變化的速率就是決定一個音素邊界的重要線索，回顧一些文獻如 Rabiner[6] 使用頻譜轉換量測 (Spectral transition measure) 的音素端點偵測方法，應用在 TIMIT 語料庫其效能可達到在誤差 20ms 的容忍範圍內，只有 15% 的音素端點位置為偵測漏失 (Missed Detection rate, MD)、22.0% 誤報率 (False Alarm rate, FA)。Kotropoulos[7] 結合 Kullback-Leibler (KL) 距離及貝式資訊法則 (Bayesian Information Criterion, BIC) 所提出的 DISTBIC 演算法來偵測語音信號之音素邊界端點，其效能在 NTIMIT 語料庫亦可達到 25.7% MD 與 23.3% FA 的結果。

在先前的語音分段或是端點偵測的研究，無論 model-based 或 metric-based 的方法中，常用的語音信號參數多與信號頻譜相關；這些參數描述了發音特徵使得語音信號的特性不同，且一般假設語音信號在短時間內為穩定的特性，故使用音框式 (frame-based) 的聲學參數，例如梅爾倒頻譜係數 (Mel-Frequency Cepstral Coefficients, MFCCs)。然而，在做頻譜分析時會造成時間與頻譜 (time-spectrum) 上之不確定性 (uncertain)，所以頻譜參數越精確就會犧牲時間精確度；但在音框式的架構中必須要讓頻譜解析度越精細，以提昇辨認音素能力，而發音器官變化很快的音素如爆破音，其音長可能小於一個音框，使得音框式的方法之語音分段位置與實際正確端點位置之間產生誤差，因此對於音素端點偵測及自動語音分段之研究來說，提昇時間解析度，必可降低大量由音框之時間解析度所造成的誤差。

由 C-H. Lee[8] 所提出之新世代自動語音辨識系統，前級首先利用語音屬性及事件偵測器群，經抽取語音特徵來偵測不同時序之語音屬性與事件，提供後級系統作語音事件及語言學知識的彙整並作證據確認及決策，達到語音辨識之目的。而在前級處理中，音素層級之邊界端點即為語音辨識系統中有用的時序資訊，語言學家亦可由這些音素端點或是時間資訊找到許多聲學特徵，提昇辨別不同音素之能力。



## 1.4 章節概要說明

本論文的内容共分為六章：

第一章：緒論：介紹本論文之研究動機與研究方向。

第二章：語料庫介紹：介紹本研究所使用之語料庫及其特性與統計分析。

第三章：取樣點式之語音聲學參數：建構取樣點為基礎的語音聲學參數。

第四章：音素端點偵測器架構：建構音素端點偵測器架構並說明其音素端點偵測器訓練之演算法。

第五章：實驗結果：對不同語料庫之音素端點偵測及自動語音分段結果進行分析，並與傳統方法比較實驗結果探討其差異。

第六章：結論與未來展望。



## 第二章 語料庫介紹

本論文將以不同語言之語料庫進行音素端點偵測或是自動語音分段的實驗，以下將對此四種語料庫作簡短介紹。在 2.1 節將介紹 TIMIT 語料庫之資料格式以及此語料庫語料中語言學上或聲學上之統計資料；在 2.2 節將介紹 TCC-300 語料庫之資料格式；在 2.3 節將介紹 Treebank 語料庫之資料格式；在 2.4 節將介紹客家話語料庫之資料格式。

### 2.1 TIMIT 語料庫簡介

#### 2.1.1 語音資料

本論文以 TIMIT[9] (The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, TIMIT) 語料庫作為主要實驗與分析之語料。TIMIT 語料庫是屬於由朗讀句子的語音 (read speech) 所組成。而語料庫中的這些朗讀語句皆是由德州儀器 (Texas Instruments, TI)、麻省理工學院 (Massachusetts Institute of Technology, MIT) 以及史丹佛研究機構 (Stanford Research Institute, SRI) 來共同設計而成。其語料庫的語句是德州儀器請美國不同區域的人朗讀並錄製成語音檔案，麻省理工學院進行人工轉寫的步驟。

TIMIT 語料庫中其包含有 6300 個語句，這些語句分別是由美國主要八種不同口音 (Dialect) 地區的 630 個語者，每位語者朗讀 10 個語句錄製而成。語料庫中其詳細的男女、地區分佈如表 2.1。語料庫語句在收錄時以 16kHz 之取樣率經 16 位元量化來錄製單聲道音檔，音檔檔頭為 1024 位元組 (byte)，以提供語音辨識為主要應用。

每個語者朗讀的 10 個語句中之語句類型，如表 2.2，包含 2 句方言 (SA) 語句，為了顯現不同地區語者口音之差異；5 句 phonetically-compact (SX) 語句，為了每個音素出現之頻率能夠相近；以及 3 句 phonetically-diverse (SI) 語句，其語句是從當時現存的文字語料庫資源挑出來的，如布朗文字語料庫 (Brown Corpus, Kuchera and Francis, 1967) 等等。



表 2.1：方言之人數分布

方言之人數分布		男性		女性		總計	
編號	區域	人數	百分比	人數	百分比	人數	百分比
1	New England	31	63%	18	27%	49	8%
2	Northern	71	70%	31	30%	102	16%
3	North Midland	79	67%	23	23%	102	16%
4	South Midland	69	69%	31	31%	100	16%
5	Southern	62	63%	36	37%	98	16%
6	New York City	30	65%	16	35%	46	7%
7	Western	74	74%	26	26%	100	16%
8	Army Brat (moved around)	22	67%	11	33%	33	5%
總計		438	70%	192	30%	630	100%

表 2.2：TIMIT 語料庫語句於不同語句類型之分布

語句類型	語句數目	語者數目	總計	每位語者之語句數目
Dialect (SA)	2	630	1260	2
Compact(SX)	450	7	3150	5
Diverse(SI)	1890	1	1890	3
總計	2342	---	6300	10

## 2.1.2 文字轉寫之人為時間標記

TIMIT 語料庫廣泛地用於各方面有關之語音研究，其原因在於語料庫內之資訊囊括完整的文字轉寫及對應不同層級之人為時間標記；文字轉寫以及其對應字詞層級 (word level) 及音素 (phone level) 的人為時間標記，使得 TIMIT 語料庫成為一個平台來提供各式各樣之理論及方法之間進行語音相關研究，並基於此平台驗證其理論、方法或是評量實驗結果效能的好壞。

無論是在何種層級之文字轉寫中，皆是由標音員給予該語音信號正確的標音符號並依其語音段落之起始與結束的語音取樣點作為時間標記，如圖 2.1 所示。如前一章節所述，文字

轉寫中的人為時間標記是目前最為準確對語音進行分段的方式，但其標記位置皆含有主觀的判斷且因人而異，容易造成時間標記之不一致性。因此將在本論文實驗分析時，來討論此現象引起的相關問題。

目前語料庫之音素集包含 61 個音素，如附錄一，音素層級之文字轉寫皆是對應音素集標記而成。但是以音素端點偵測的觀點觀察語音信號的變動時，不同音素語音信號之轉變其無論在頻域或是時域上之特性應是有所差異的，利用此差異我們可以偵測音素端點存在的可能性。而在爆破音（stop consonant）發音前會有所謂短停頓的產生，在語音學上稱為嗓音起始時間（voice onset time, VOT），指的是爆破音成阻後持阻到除阻時間，語音學上會將此段短停頓的產生視為爆破音時長的一部份。但在音素端點的偵測內，其語音信號的特性上卻是有著極大的差異。故 TIMIT 語料庫的音素時間標記將此種情形也納入音素時間標記的範疇中，而對該爆破音之標音前的短停頓給予合適的標記符號，其對應的標記符號如下表 2.3。

另外，我們知道英語為 consonant-vowel-consonant 之音節結構，簡稱為 CVC。例如以（rime structure）表示單音節的英文詞 cat，其音節頭（onset）為“c”，音節核為“a”，音節尾（coda）為“t”。而子音在 CVC 音節結構內的位置不同會其發音也不盡相同，以本論文之音素端點偵測的觀點，我們無須了解其音素在結構內的關係，但若以音素端點切割的方面考量，就必須考慮音節結構對音素端點的影響。

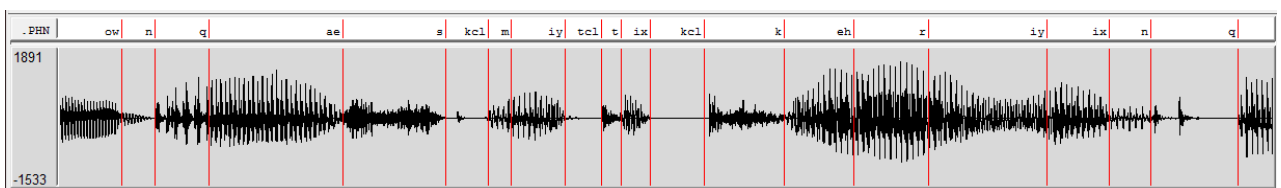


圖 2.1：音素層級之文字轉寫對應於語音信號的人為時間標記

表 2.3：爆破音對應之短停頓標記符號。

stops	b	d	g	p	t	k	jh	ch
closure intervals	bcl	dcl	gcl	pcl	tcl	kcl	dcl	tcl

TIMIT 語料庫之訓練語料與測試語料分別為 462 位語者之 4620 個語句與 168 位語者 1680 個語句所建構而成，在本論文中使用音素層級之文字轉寫的人為時間標記之所有訓練語料來訓練音素端點偵測器的模型，並以測試語料來提供給本論文所提出方法之實驗。以下為 TIMIT 語料庫之檔案結構說明：

CORPUS ::= timit	/*語料庫名稱*/
USAGE ::= train   test	/*訓練與測試語料*/
DIALECT ::= dr1   dr2   dr3   dr4   dr5   dr6   dr7   dr8	/*不同口音之區域分類*/
(如表 2.1 方言人數分布之區域編號)	
SEX ::= m   f	/*語者性別*/
SPEAKER_ID ::= <INITIALS><DIGIT>	/*語者資料名稱命名*/
INITIALS ::= speaker initials, 3 letters	
DIGIT ::= number 0-9 to differentiate speakers with identical Initials	
SENTENCE_ID ::= <TEXT_TYPE><SENTENCE_NUMBER>	/*語句名稱命名*/
TEXT_TYPE ::= sa   si   sx	/*語句類型*/
(如表 2.2 不同語句類型之分布)	
SENTENCE_NUMBER ::= 1 ... 2342	/*語句編號*/
FILE_TYPE ::= wav   txt   wrd   phn	/*檔案類型*/
(依序為音檔、語句文字、字詞時間標記、音素時間標記)	

## 2.2 國語 TCC-300 語料庫簡介

本論文中使用 TCC-300 麥克風語音資料庫是由國立交通大學、國立成功大學、國立台灣大學所共同錄製，中華民國計算語言學學會所發行，此語料庫屬於麥克風朗讀語音，主要目的是為提供語音辨認研究，檔案統計資料如表 2.4 所示。台灣大學語料庫主要包含詞以及短句，文字經過設計，考慮音節與其相連出現之機率，共 100 人，每人錄製一句而成；成功大學及交通大學為長文語料，其語句內容由中研院提供之 500 萬詞詞類標示語料庫中選取，每篇文章包含數百個字，再切割成 3 至 4 段，每段至多 231 字，分別各 100 人，每人錄製一句朗讀來錄製，且每人所朗讀之文章皆不相同。每個學校之語句取樣頻率皆為 16000 赫茲 (Hertz)，取樣位元數為 16 位元。音檔檔頭為 4096 位元組 (byte)，副檔名為\*.vat。

表 2.4：TCC-300 語料庫檔案統計資料

學校	語音檔案(*.vat)	文字檔案(*.tab)	群集(Group)
交通大學	1238	1238	5
成功大學	1170	1170	5
台灣大學	6509	6509	1

屬於聲調語言之國語音節結構如下圖所示可將音節分為聲母和韻母，韻母可再細分為介音與韻腳，而韻腳包含主要元音和韻尾，而本論文使用之 TCC-300 國語語料庫是以類音素單元做為自動語音標記的基本語音單元，類音素即是將國語結構分為聲母、韻母（但韻母不包含鼻音韻尾）以及鼻音韻尾等三個部份以依照語音之特性簡化結構。

在 TCC-300 語音資料庫之語料選取方面，我們使用交通大學與成功大學所錄製的長文語料，並隨機選取六分之五的部份當作訓練語料，其它部分為測試語料。本論文提出自動標記音素位置之方法是以兩個階段 (two-stage) 來達成自動語音分段的目標，故需要有一個初始位置來訓練一個自動端點偵測器，以進行第二階段更進一步地修正。由於 TCC-300 語音資料庫沒有人工標記的音素分段位置，利用 HTK (Hidden Markov Toolkit) 使用 SAT (speaker adaptation transform, feature MLLR) 及 SA (speaker adaptation, MLLR) 技術訓練 HMM 類

音素模型，獲得較佳的 HMM 模型後進行強迫對齊 (force alignment) 之自動分段結果，作為 TCC-300 語料庫之類音素初始分段位置，以提供本論文使用。

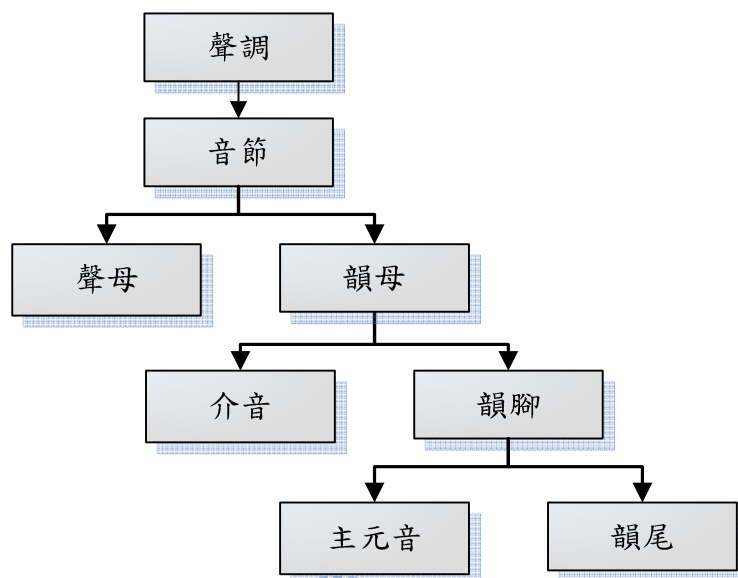


圖 2.2：國語音節結構圖

## 2.3 國語 Treebank 語料庫簡介

Treebank 語料庫包含 425 個語句且含有 56237 個音節，由一個專業的女性播音員所錄製。此語料庫屬於麥克風朗讀語音，主要目的是為提供語音韻律標記與建立韻律模型之研究。語句取樣頻率皆為 16000 赫茲 (Hertz)，取樣位元數為 16 位元，副檔名為 \*.pcm。

在文字轉寫方面，因 Treebank 語料庫內含人為時間標記之音節與聲/韻母層級文字轉寫，本論文以此兩種層級之文字轉寫作為實驗結果之標準答案以評量實驗結果之效能。另外，藉由 HTK toolkit 訓練音節以及聲/韻母 (initial/final) 語音單元之 HMM 模型，對語料庫進行強迫對齊，獲得初始自動分段位置用於實驗使用與測試。選擇梅爾倒頻譜係數作為語音聲學參數，參數設定為 38 維，其中包含 12 階的梅爾倒頻譜係數與能量之對數值 (log energy) 及其一階差量與二階差量並扣除原本的能量對數值總計 38 維，音框長度 (frame length) 設為 32 毫秒，音框平移 (frame shift) 設為 5 毫秒。

Treebank 語料庫在訓練及測試語料的選擇上，扣除語句中含有英文的 4 個語句，剩下 421 句以 9：1 的比例隨機選取，得訓練語料為 379 句和測試語料為 42 句。

## 2.4 客語語料庫簡介

本論文為使用四縣客家話語料庫，文章出處為龔萬灶老師所撰寫的「阿啾箭个故鄉」，音檔取樣頻率為皆以 20k 赫茲及取樣位元數為 16 位元之單聲道錄製而成，副檔名為\*.pcm 格式。語料庫之語者為龔老師共錄製語音檔案 639 個，包含 42 篇文章共有 63158 個音節。語音檔是由發音人在普通房間依照文稿唸出，屬於朗讀式語音並依照錄製之日期、文章編號來命名。

在文字轉寫方面，因客語音節結構與國語相同，在此本論文以聲/韻母作為語料庫的文字轉寫之基本單元，而客語語料庫無人為時間標示之音素端點位置可提供正確的端點進行訓練。藉由 HTK 訓練聲/韻母之 HMM 模型，對語料庫進行強迫對齊以獲得四縣客語文字轉寫之初始自動分段位置。使用梅爾倒頻譜係數做為聲學參數，參數設定為 38 維，其中包含 12 階的梅爾倒頻譜係數與能量之對數值及其一階差量與二階差量並扣除原本的能量對數值總計 38 維，音框長度設為 32 毫秒，音框平移設為 5 毫秒。

客語語料庫在訓練及測試語料的選擇上，同樣以 9：1 的比例隨機選取，訓練語料為 587 句和測試語料為 73 句。

## 第三章 取樣點式之語音聲學參數

傳統聲學參數與本論文所提出之取樣式聲學參數最大的差異即是時間與頻譜的取捨，在傳統上抽取聲學參數方式通常假設語音信號為短時間穩定而依固定的取樣點數作為一個音框，音框可視需要改變音框平移以及音框長度，並以此音框為單位抽取語音信號的聲學參數。音框平移的寬度影響時域上音素標記的精準度，音框長度影響著語音信號在頻譜之細膩程度。但在音素分段的觀點，上述這兩種影響卻是不必要的，語音信號的特性雖表現於頻譜分佈上，不過語音信號為時變的，音框式之時間解析度較低，音素之端點位置即使標記在正確的音框內仍會與實際正確端點位置之間產生誤差。本論文所使用的聲學參數結合語言學家所提出的聲學參數，並應用於本論文所提出之音素端點偵測以及自動音素分段的研究方法。3.1 節將介紹所提出之取樣點式聲學參數之語音特徵特性；3.2 節為利用取樣點式聲學參數之特性來進行類音素端點自動分段之初步實驗結果。

### 3.1 取樣點式聲學參數之語音特徵

本論文提出一些取樣點式聲學參數如子頻段之信號波封[10] (sub-band signal envelope)、上升率[10](rate of rise, ROR)、頻譜熵[11](spectral entropy)、頻譜 KL 距離(spectral KL distance)，列舉數個聲學參數範例以觀察在不同語音信號或是語音屬性的變化時呈現出的聲學特性為何。以下，進一步介紹本研究所使用的語音特徵參數：

#### 3.1.1 子頻段信號波封

在語言學家所提出的聲學參數中，有許多帶通濾波器能量 (band-energy)，它們各自能用來區別不同的發音方式或發音位置，常見的頻段[10] (filter bank) 有以下：

0.0 – 0.4 kHz 0.8 – 1.5 kHz 1.2 – 2.0 kHz

2.0 – 3.5 kHz 3.5 – 5.0 kHz 5.0 – 8.0 kHz



例如在摩擦音、塞擦音中，在頻譜中之高頻段成份能量極強，低頻段成份能量較弱，鼻音韻尾或是母音的部分則是在低頻段的成份能量極強。這些頻段中能量在有明顯變化的時候，可視為是語音信號開始改變的地方。但語言學家所使用的聲學參數為信號波封 (signal envelope)，而非現今語音辨認器中常用的能量。故我們將這六個頻段能量取出它的波封來當作本研究中所使用的聲學參數。

在製作一個波封檢測器 (envelope detector) 的同時，為了保持在波封變化時之信號能正確地描述信號的波封變化，其變化即為頻段信號波封的表示方式；使用希爾伯特變換 (Hilbert transform) 來求取輸入信號的波封是一個適當且普遍的方法，其中  $H(x[n])$  為輸入信號  $x[n]$  的希爾伯特變換，若輸入信號為頻段之能量  $x[n]$ ，其  $H(x[n])$  即為語言學家所使用信號波封，如下式：

$$H(x[n]) = x[n] \otimes h[n] \quad \text{and} \quad h[n] = \begin{cases} 0, & n \text{ is even} \\ 1/n\pi, & n \text{ is odd} \end{cases} \quad (3-1)$$

圖 3.1 即為語音信號經波封檢測器輸出之波封結果，其表示語音信號的輪廓，但是觀察輪廓時卻沒有明顯的規則可做為分辨音素端點的依據，故轉而觀察語音信號在使用六個頻段中之分佈，並依此分佈之特性來區分不同的音素。

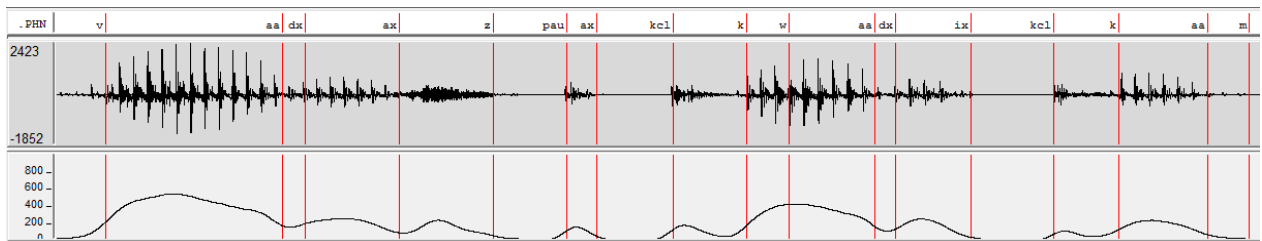


圖 3.1：取樣式語音波封聲學參數範例，由上至下分別表示音素層級之人為時間標記的文字轉寫、語音信號、音高軌跡 (pitch contour)、語音信號之波封

另外，考慮語音信號之波封受到喉頭震動的影響 (尤其在音高較低的男性影響越顯著)，其會造成語音信號的特性與喉頭震動的周期產生某種程度的關聯性或是造成語音信號的不連貫性，使得波封出現不是預期該有的波動而產生失真。為避免如以上所述之影響，藉由調



整波封檢測器的低通濾波器頻寬 (passband bandwidth)、截止頻率的衰減斜率<sup>1</sup>來達到其參數物理意義之目的。由簡單的頻寬-濾波器階數定性分析發現，低通濾波器頻寬在 30Hz 至 50Hz 之間並使用相同之濾波器階數，其語音信號波封的輸出結果沒有太大的差異，但其波封變動卻與不同之濾波器階數影響最大，圖 3.2 即是顯現出以上所述之觀察結果。

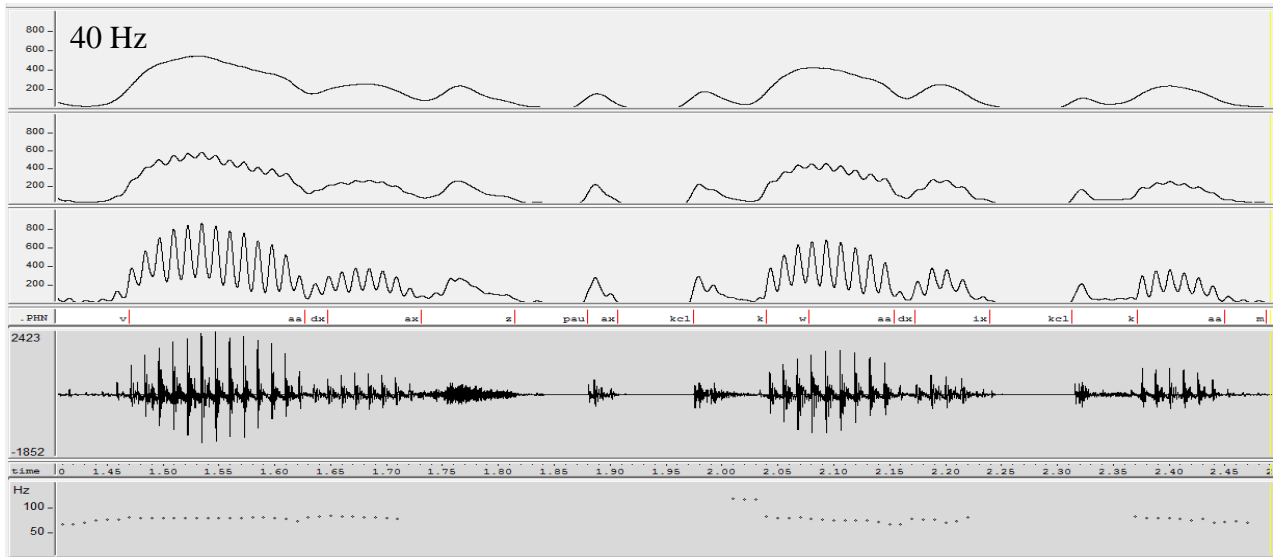


圖 3.2：不同階數之波封檢測器輸出結果，由上至下分別表示波封檢測器使用 40Hz 之 641 階、321 階、161 階低通濾波器的語音信號波封輸出結果、音素層級之人為時間標記的文字

轉寫、語音信號、音高軌跡

<sup>1</sup>濾波器的衰減斜率是指由濾波器之頻率特性曲線上，從濾波器的截止頻率或是衰減的曲線之頻率每提高或下降一個倍頻其信號衰減的分貝值。

### 3.1.2 上升率

語言學家所稱之上升率，可用於描述語音信號之聲學參數變動的情況，因此藉由參數變動量而觀察發現可能存在的音素端點；其計算方法可對應於音框式抽取語音特徵參數的一階時間軸導數（time derivatives）的方式：在有限的視窗寬度（window width）內，第  $n$  個取樣點的上升率  $ROR_x[n]$  依據對應的輸入參數所計算如下式：

$$ROR_x[n] = \frac{\sum_{i=-w}^w i \cdot x[n+i]}{\sum_{i=-w}^w i^2} \quad (3-2)$$

其中  $x[n+i]$  為輸入參數資料， $w$  為計算上升率所使用的視窗寬度。本研究使用語音波形之波封的上升率、頻譜熵之上升率、各頻段信號波封的上升率等當作語音信號的聲學參數，來評量各取樣點式聲學參數的變化率。

透過觀察下圖 3.3 可以發現由人為時間標記對應於語音信號之波封急遽上升的時候，即是該區域波封上升率之局部最大值（local maximum）之端點。在此處之上升率參數可指出語音信號之波封變動最大的端點位置，這種情況尤其好發在音節結構的前端音節頭至音節核的部分，如摩擦音至母音、塞擦音至母音…等等的音素轉換端點，由以上觀察的聲學參數之特性，我們將其輸入參數至換成各頻段的信號波封，那麼我們即可由各頻段信號波封所計算的上升率來分別找到對應每個頻段其信號波封變動量大的端點。如圖 3.4 各頻段的波封上升率可以對應於聲譜圖<sup>2</sup>（spectrogram）的顏色深淺程度，也就對應至各頻段信號波封的大小變化；語音信號在六個頻段之中之分佈由亮轉灰暗，其轉變程度越大上升率越高。然而，觀察每個頻段之波封上升率為局部最大值之端點，其會因為信號波封變動量的不同而使得在某一段時間內各頻段之端點位置並不一致，要如何在此一區段時間選擇一個適當的音素轉換端點，將在下節討論。

---

<sup>2</sup>聲譜圖是以 2 維影像來呈現時變的語音信號在頻譜上的分佈以及強度，常被用以分析不同音素之語音特性。

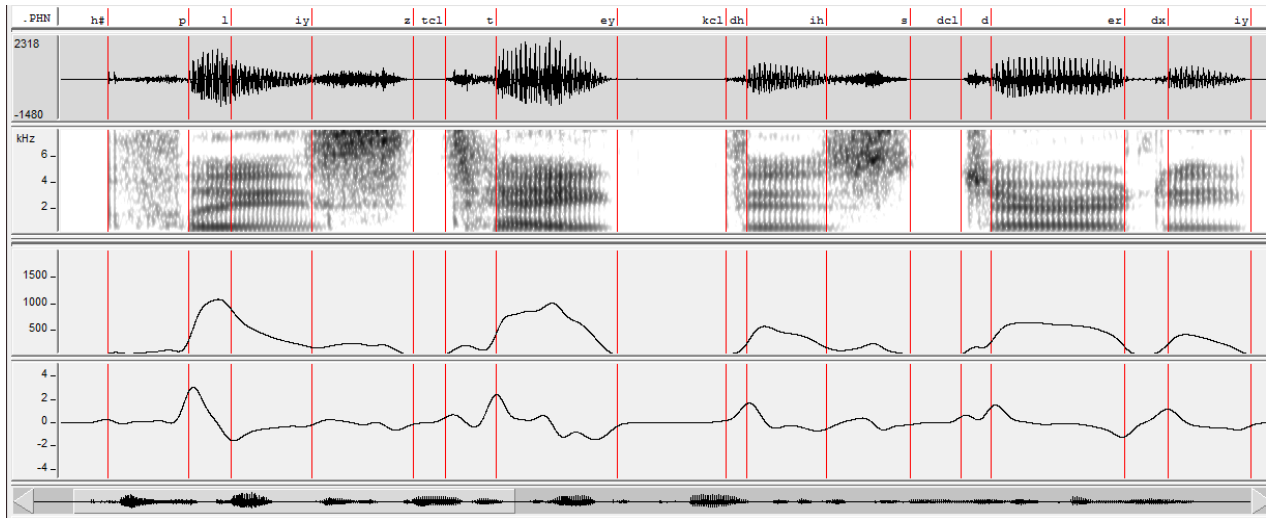


圖 3.3：取樣式聲學參數之上升率範例，由上至下分別表示音素層級之人為時間標記的文字轉寫、語音信號、聲譜圖、語音信號之波封、波封之上升率

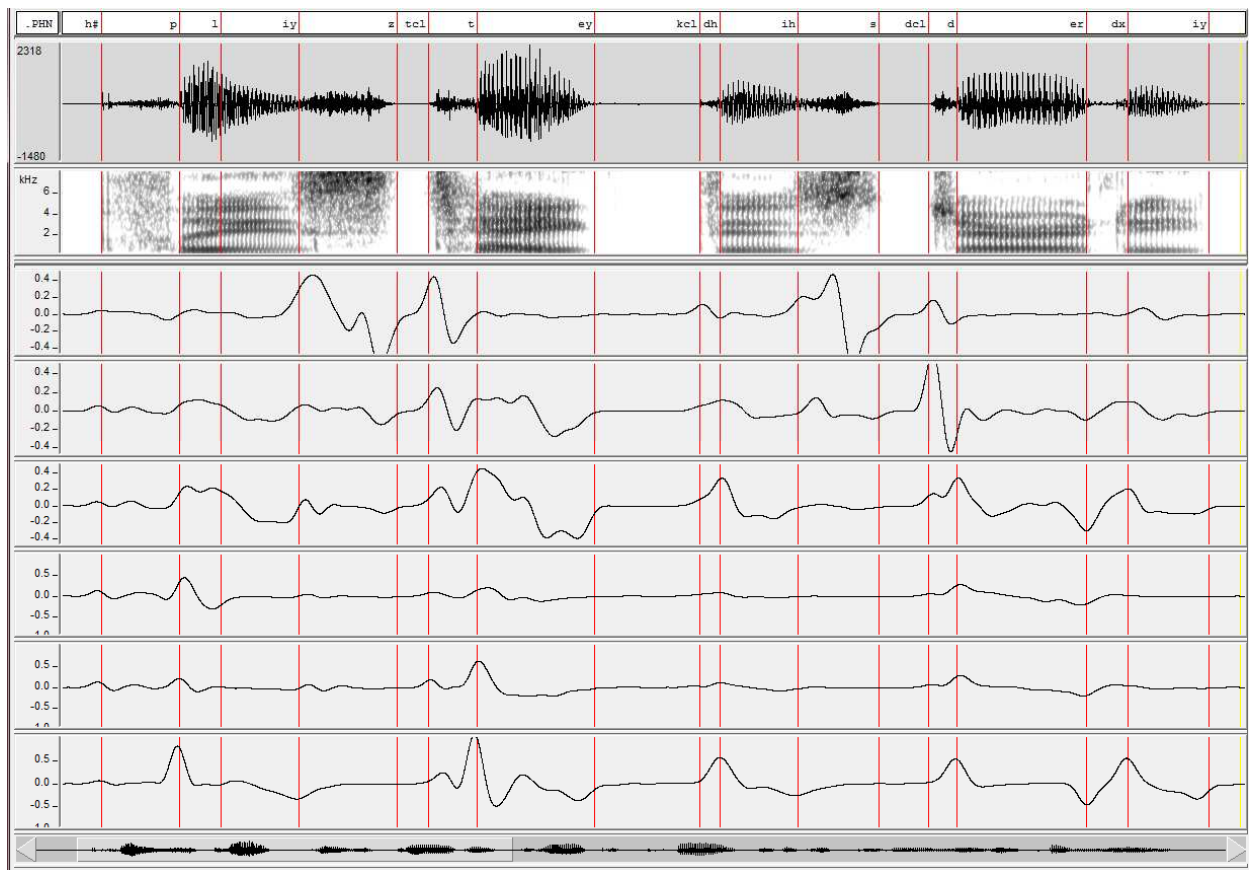


圖 3.4：取樣式子頻段信號波封聲學參數範例，由上至下分別表示音素層級之人為時間標記的文字轉寫、語音信號、聲譜圖、第六個至第一個頻段之信號波封上升率

### 3.1.3 頻譜熵

頻譜熵可用來描述信號在頻譜上的集中之分佈程度，若語音信號越集中在某一個頻段則頻譜熵越小。在此，本研究使用先前所述之六個頻段，將頻譜熵  $H_s[n]$  定義如下式表示：

$$H_s[n] = -\sum_i E_i[n] \log(E_i[n]) \quad (3-3)$$

$$\text{其中 } E_i[n] = \frac{e_i}{\sum_{j=1}^6 e_j} \quad (3-4)$$

$E_i[n]$  為第  $i$  個頻段之第  $n$  點正規化之後的子頻段信號波封。由語音信號對應到頻譜熵的表現上如圖 3.5，可以發現短停頓、靜音內之語音特性只有非語音的雜訊。如背景雜訊在各個頻段都會出現，所以頻譜熵值較高是可以預期的；而母音在頻譜上的能量則較集中於低頻段至中頻段的部分，其頻譜熵值相對較低。同樣地，可依頻譜熵在不同之音素在頻譜上的分佈之間的變動，求取頻譜熵的上升率。

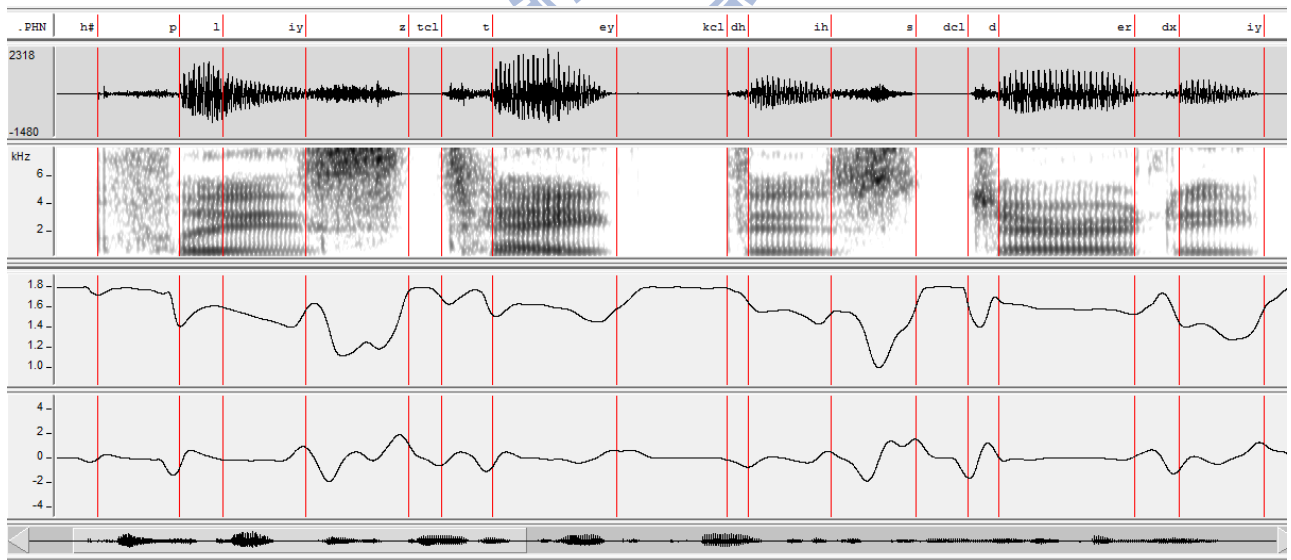


圖 3.5：取樣式頻譜熵聲學參數範例，由上至下分別表示音素層級之人為時間標記的文字轉寫、語音信號、聲譜圖、頻譜熵、頻譜熵之上升率

### 3.1.3 頻譜 KL 距離

將頻譜視為一個機率分佈的問題，因此可以利用頻譜 KL 距離來描述兩段時間點之頻譜相似程度。在語音信號中計算兩點不同時間( $n$  與  $m$ )的頻譜 KL 距離， $d_{KL}(n,m)$ ，可以由下式表示：

$$d_{KL}[n,m] = \sum_{i=1}^6 (E_i[n] - E_i[m]) \log \left( \frac{E_i[n]}{E_i[m]} \right) \quad (3-5)$$

而本研究目前為考慮**相鄰語音信號取樣點**之頻譜信號分佈特性，則將(3-5)式改寫為以下：

$$d_{KL}[n] = \sum_{i=1}^6 (E_i[n] - E_i[n+1]) \log \left( \frac{E_i[n]}{E_i[n+1]} \right) \quad (3-6)$$

不同音素轉換的時候，其發音的方法或是部位也會跟著轉移，使得不同音素之語音信號轉換至頻譜上的分布情形也會跟著不同，頻譜 KL 距離即是度量在頻譜間的相似程度，且此一度量之特性具有一致性。那麼經由簡單調整一個臨限值 (threshold)，即可初步地得到一序列 (sequence) 經由頻譜 KL 距離所挑選出來是具有音素端點可能性的位置。

藉由聲譜圖可以清楚地觀察到在相鄰音素之間的信號分佈變化，如圖 3.6 中同一音素內之頻譜信號分佈為局部穩定的狀態，並在不同音素轉換的區域因其頻譜分佈差異大，使頻譜 KL 距離明顯增大。

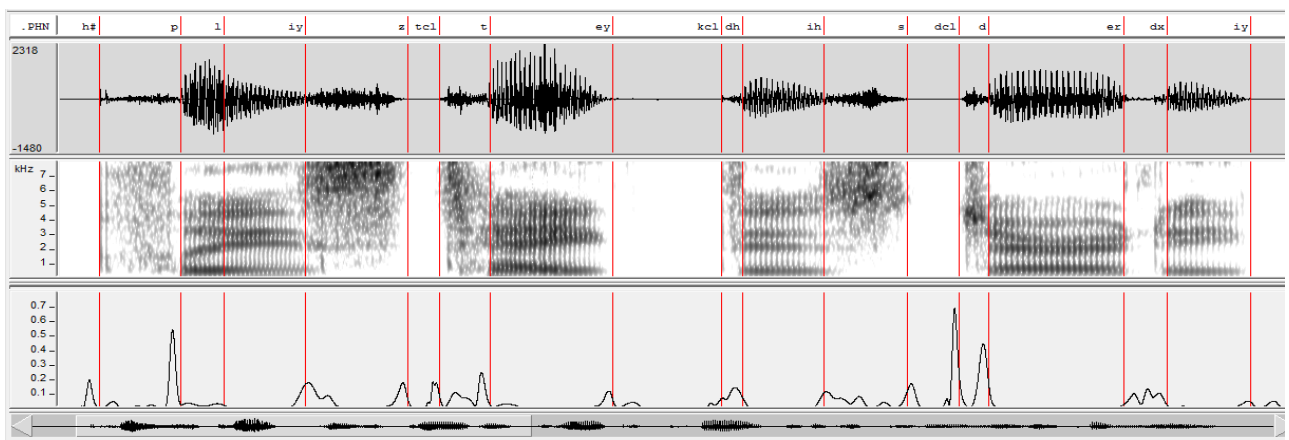


圖 3.6：取樣式頻譜 KL 距離聲學參數範例，由上至下分別表示音素層級之人為時間標記的文字轉寫、語音信號、聲譜圖、頻譜 KL 距離



由前 3.1.1 節所述波封檢測器內低通濾波器之階數，影響著頻段內之信號波封的變動。利用各頻段分佈所計算出來的頻譜 KL 距離也有如下圖 3.7 的差異，在圖中我可以觀察到隨著濾波器階數越低，則頻譜 KL 距離的大小因信號變化而受影響的程度也會增加。假若使用臨限值來挑選一序列音素之候選端點，在高階數的部分，音素端點之候選端點少，其端點雖能表現出信號的重大變化，但有部分的音素端點卻因為臨限值之遮蔽而消失；相對地在低階數的部分，情況卻是完全相反，序列中音素候選端點幾乎能包含原有之音素端點，不過因為其頻譜 KL 距離易受信號變化影響的效應，使得音素候選端點序列中增加極多冗餘的端點。那麼以音素端點偵測的觀點考量，就必須在音素候選端點的數目與參數的穩定度上做一個取捨 (trade-off)，以達到最佳的結果。

綜合以上所敘述之取樣點式聲學參數，其子頻段信號波封、聲學參數的上升率、頻譜熵及頻譜 KL 距離等語音特徵參數的變化，確實能得到在語音信號變化的時候，可以觀察這些參數的語音特性達到分辨不同音素端點位置之目的。

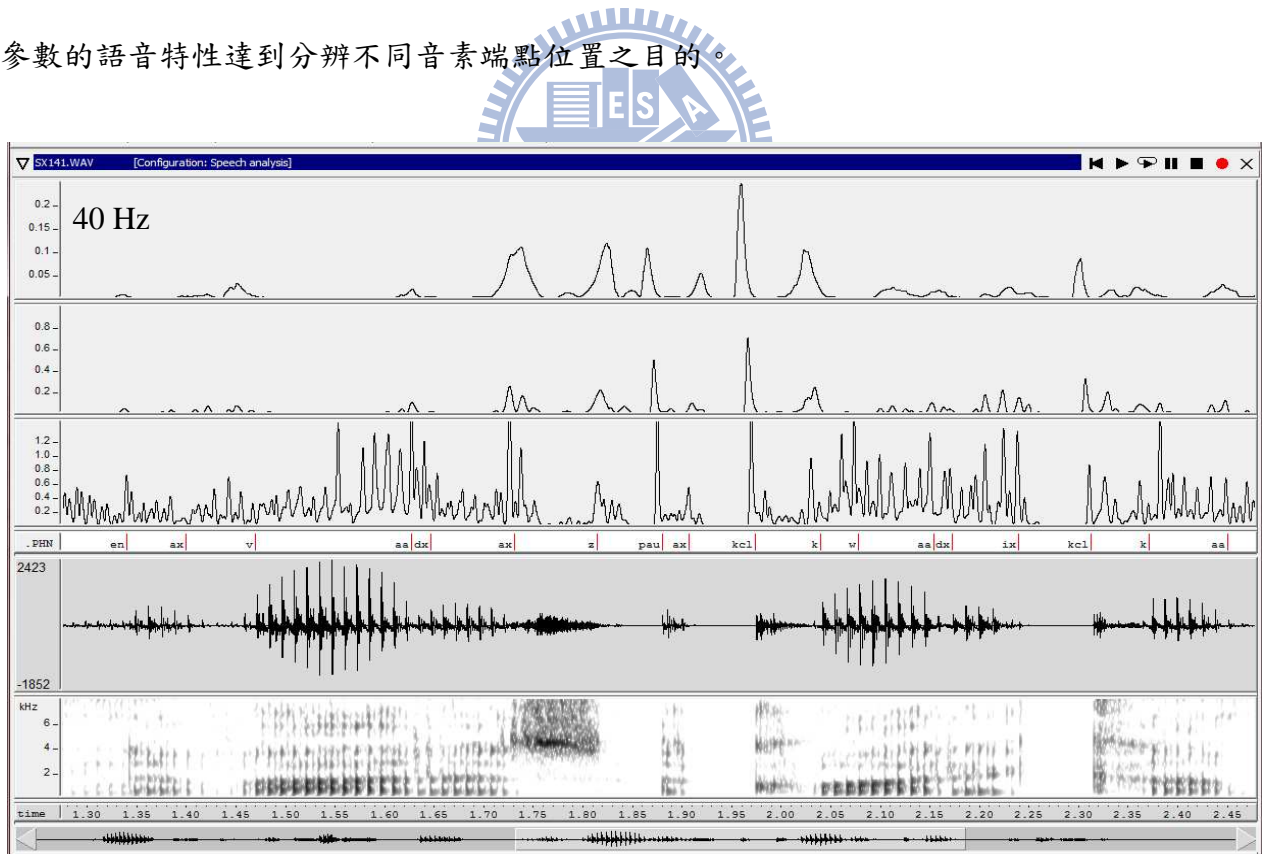


圖 3.7：不同階數之波封檢測器對頻譜 KL 距離的影響，由上至下分別表示波封檢測器使用 40Hz 之 641 階、321 階、161 階低通濾波器輸出結果所計算的頻譜 KL 距離、音素層級之人為時間標記的文字轉寫、語音信號、聲譜圖

## 3.2 使用取樣點式聲學參數之類音素端點自動分段

由於國內有人工的正確標記語音位置之國語語料庫不多，而不正確的標音位置會造成後級辨識系統或是合成系統之效能無法提昇。因此，一個使用自動且分段位置精準的方法，可以大幅提昇後級的效能。過去的研究常以音框式之 HMM 架構為基礎來獲得之語音標記位置作為正確標示，此方法雖然可以達成自動語音分段的目的，但最終還是需要人工進一步修正，離正確語音的標記位置之間仍有許多改進的空間。本論文以 3.1 節所提出之取樣式聲學參數之特性，對國語語料庫 TCC-300 進行自動分段的初步實驗，其步驟如下：

首先，利用 SAT (Speaker Adaptation Transform, feature MLLR) 及 SA (Speaker Adaptation, MLLR) 後的出語者調適 HMM 模型來做 TCC-300 的類音素單元之初始自動語音分段位置，接著利用此初始位置依照發音方法的不同做分類，如表 3.1。並由初始位置當作參考位置再利用取樣式聲學參數的特性來調整音素端點之標記位置。以下比較 HMM 之初始位置及以取樣式聲學參數特性修正後之語音分段位置。

表 3.1：國語語音發音方法的分類表。

發音方法(Manner)	發音方法對應之音素					
爆破音 Stop	b	p	d	t	g	k
鼻音 Nasal	m	n	(n_n)	(ng)		
摩擦音 Fricative	f	s	x	h	sh	
塞擦音 Affricate	q	j	c	z	zh	ch
流音 Liquid	l	r				
韻母音 Vowel	others					

先前在觀察 HMM 自動語音分段位置的準確度時，發現短停頓常會有無法標記出來或是標記位置錯誤之情形，而使得某些音素之平均音長有過長的現象，如塞擦音與爆破音等。在此本論文使用信號波封與各頻段之信號波封來判斷語音段是否為短停頓的狀態。由圖 3.8 可

以觀察到短停頓中各個頻段之信號波封與其它有語音信號的地方相比其數值幾乎非常地低且根據語音屬性不同而有不同的頻譜分佈情形。在此，簡單以信號波封與各頻段之信號波封來標記短停頓的端點。短停頓標記修正之演算法如下：

- (1) 前端點：在原端點位置之前後 30 毫秒的範圍內，判斷語音波形之波封是否小於波封之臨限值而得到一個交集點，再經由交集點附近距離 10 毫秒內來判斷各個頻段之信號波封是否小於頻段波封之臨限值的條件作聯集來決定是否有短停頓的狀態。
- (2) 後端點：在原端點位置之前後 30 毫秒的範圍內，判斷語音波形之波封是否大於波封之臨限值而得到一個交集點，再經由交集點附近距離 10 毫秒內來判斷各個頻段之信號波封是否大於頻段波封之臨限值的條件作聯集來決定是否有短停頓的狀態。

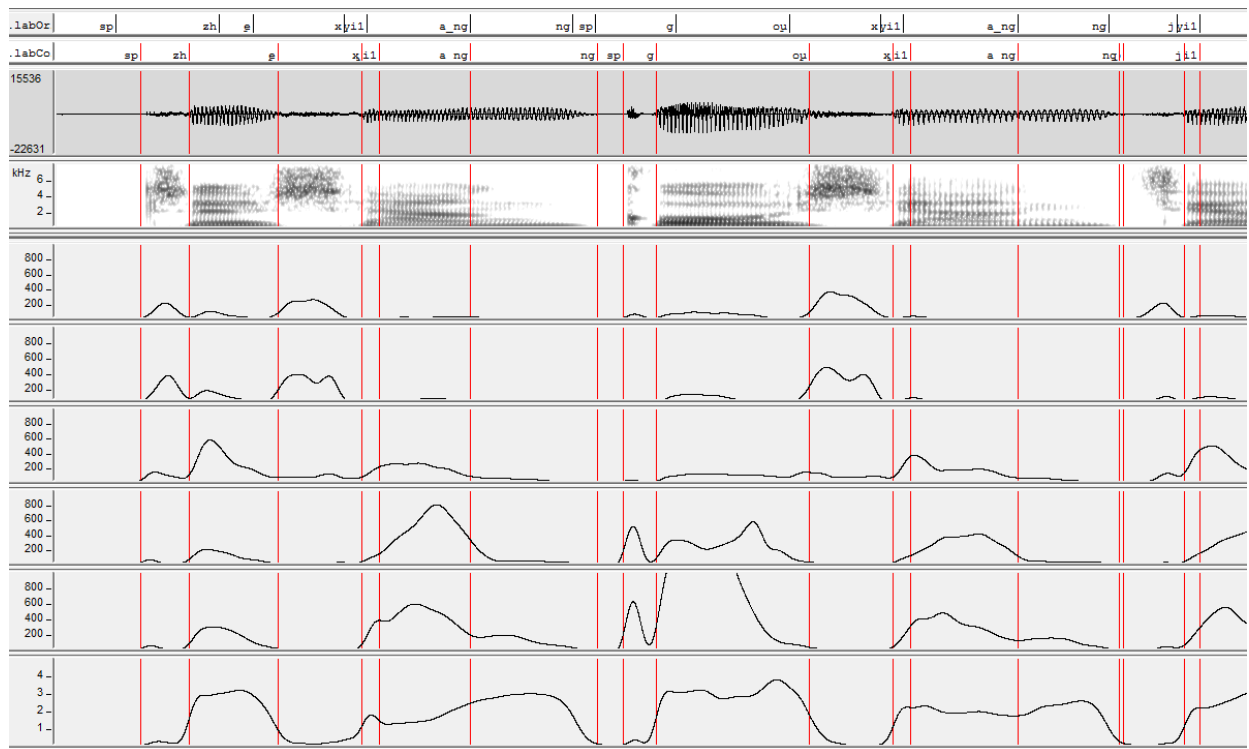


圖 3.8：國語語句端點位置自動調整(短停頓)演算法則之範例，最上方兩列標音位置分別表示是 HMM 自動語音分段及修正後之語音標記位置；接著由上至下的圖形分別表示語音波形、聲譜圖、第六個至第一個頻段的信號波封



接下來觀察摩擦音、塞擦音等發音方法之音素，其在於頻譜中與相鄰母音與短停頓有極大的頻譜差異。在此，使用頻譜 KL 距離、頻譜熵及頻譜熵的上升率來調整音素的端點。圖 3.9 所示，由摩擦音與塞擦音頻譜中可觀察到頻譜 KL 距離在母音轉換至摩擦音、塞擦音之間有較高的峰值，且摩擦音、塞擦音相鄰母音的端點，其頻譜熵值上升與下降速度很快，分別在頻譜熵的上升率中造成極大、極小的峰值。頻譜熵的上升率之峰值位置與人所期望的正確端點位置差距不遠，由先前研究可以了解頻譜熵、頻譜 KL 距離等已知在音框式量測信號變化量方法中是非常有用的聲學參數，同樣在取樣式聲學參數量測信號變化量的效果一樣明顯，且語音之分段位置更精準。

摩擦音、塞擦音程式修正演算法如下式：

- (1) 後端點：找到此一區段頻譜熵上升率的相對極小值，在小範圍的搜尋 KL distance 相對極大值。
- (2) 前端點：利用後端點的位置當做參考位置，判斷前面是否有短停頓，有則利用短停頓的方式偵測前端點，若無短停頓則搜尋一段範圍找到此一區段頻譜熵上升率的相對極大值。

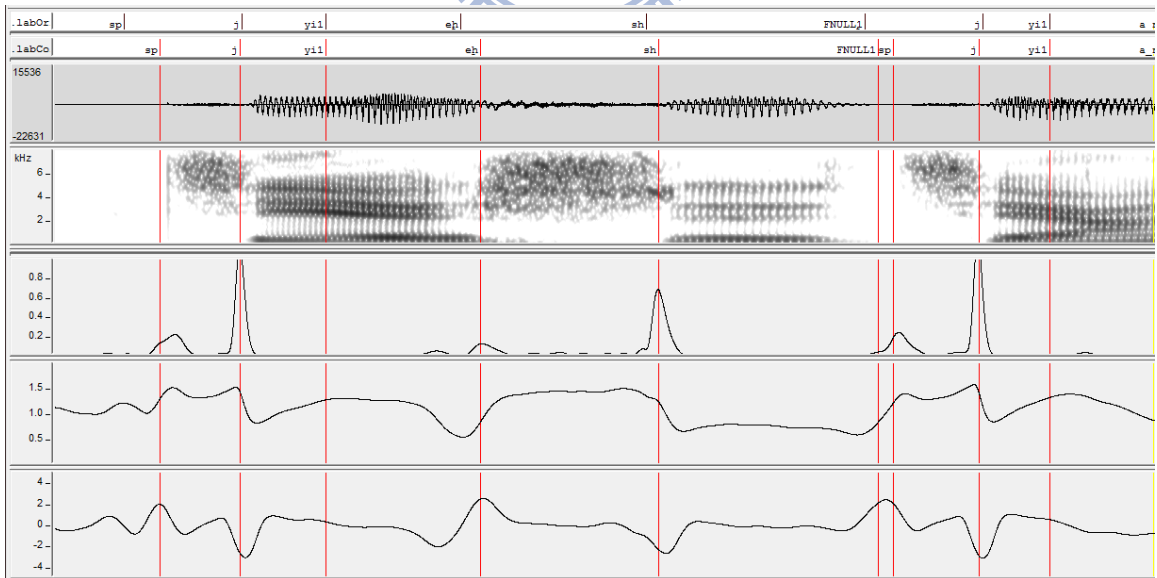


圖 3.9：國語語句端點位置自動調整(摩擦音、塞擦音)演算法則之範例，最上方兩列標音位置分別表示是 HMM 自動語音分段及修正後之語音標記位置；接著由上至下的圖形分別表示語音波形、聲譜圖、頻譜 KL 距離、頻譜熵、頻譜熵上升率

爆破音切割位置的修正時，由波形與頻譜觀察中發現通常在爆破音開始的時候會有短停頓出現，接著波封會有急遽上升的現象，故本論文使用波封之上升率來描述其現象。如圖 3.10 中(a)、(b)小圖所示，在爆破音結束的地方，也是音素轉換的端點。

爆破音程式修正演算法如下式：

- (1) 後端點：找到此一區段波封上升率的相對極大值，並在該極大值之位置找到頻譜 KL 距離的相對極大值。
- (2) 前端點：利用後端點的位置當做參考位置，判斷前面是否有短停頓，有則利用短停頓的方式偵測前端點，若無短停頓則搜尋此一區段之頻譜 KL 距離的相對極大值。

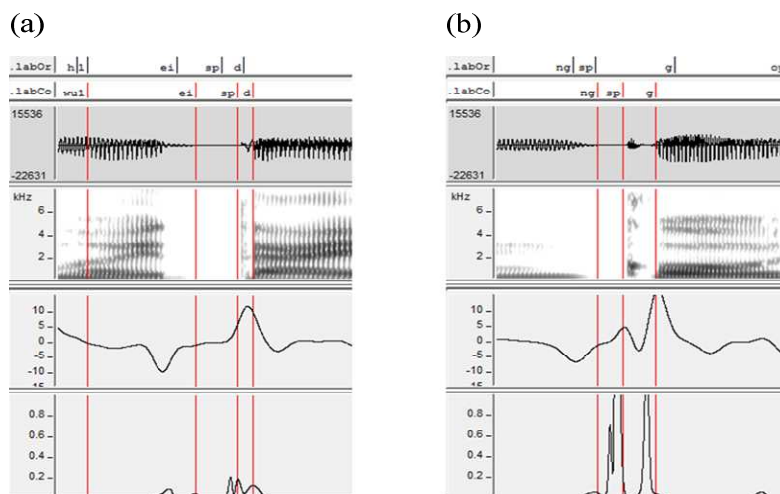


圖 3.10：國語語句端點位置自動調整(爆破音)演算法則之範例：(a) d 和(b) g 最上方兩列標音位置分別表示是 HMM 自動語音分段及修正後之語音標記位置；接著由上至下的圖形分別表示語音波形、聲譜圖、語音波封上升率、頻譜 KL 距離

另外，鼻音部分可由其語音信號之頻譜分佈多集中在 0.0 – 0.4 kHz 與 0.8 – 1.5 kHz 的低頻頻段的現象，且與相鄰的音素皆有頻譜上的差異，在此我們也使用頻譜 KL 距離來判斷。鼻音部分程式修正演算法如下式：

- (1) 後端點：由原端點位置搜尋頻譜 KL 距離大於臨限值的位置。
- (2) 前端點：利用後端點的位置當做參考位置，判斷前面是否有短停頓，有則利用短停頓的方式偵測前端點，若無短停頓則搜尋此一區段之頻譜 KL 距離的相對極大值。

最後，母音端點的偵測是利用相鄰母音、子音及短停頓之端點位置，當作母音的端點位置。由實驗觀察 3.1 節所述之聲學參數特性用於自動分段位置的準確度，並與原本 HMM 初始語音分段位置作為比較對象，以下列舉 2 個實驗結果之範例，圖 3.11 與圖 3.12。首先由圖 3.11 與 3.12 中，將實驗修正後的語音標記位置對應至語音波形及聲譜圖觀察，實驗結果在音素之端點位置皆能調整到適當的地方。以方形圈圈選處之聲譜圖中，以紅色線條為分界點，其前後兩段之語音信號分佈可明顯看出實驗結果能夠將端點位置近乎正確地標示出來，而其他標記位置之準確度也同樣有好的自動標記效能。另外，有些標記位置是與 HMM 的分段位置為相同標記位置，原因在於進行實驗的過程當中，若不符合自動調整演算法之條件，其標記位置則維持不變。

自動調整端點演算法之實驗結果顯示了使用取樣點式聲學參數之特性確實有助於尋找更佳的端點位置，但演算法所使用之規則是基於聲學參數對應語音信號的觀察與語言學知識相互組合而成。然而語音信號的變化並非有一定的規則可循，故本論文將利用類神經網路之特性將各聲學參數之特性作統計分析的彙整，來找出最佳音素端點位置。

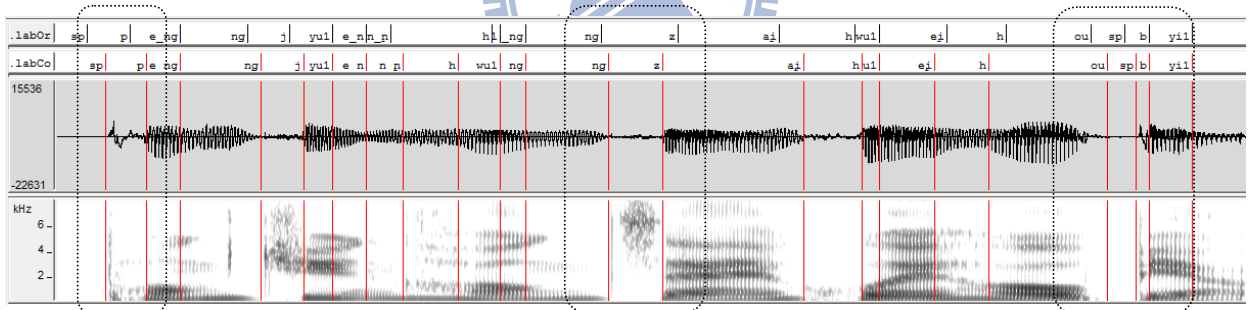


圖 3.11: 自動調整國語語句端點位置實驗結果之範例一，最上方兩列標音位置分別表示 HMM

自動語音分段及修正後之語音標記位置、語音波形、聲譜圖

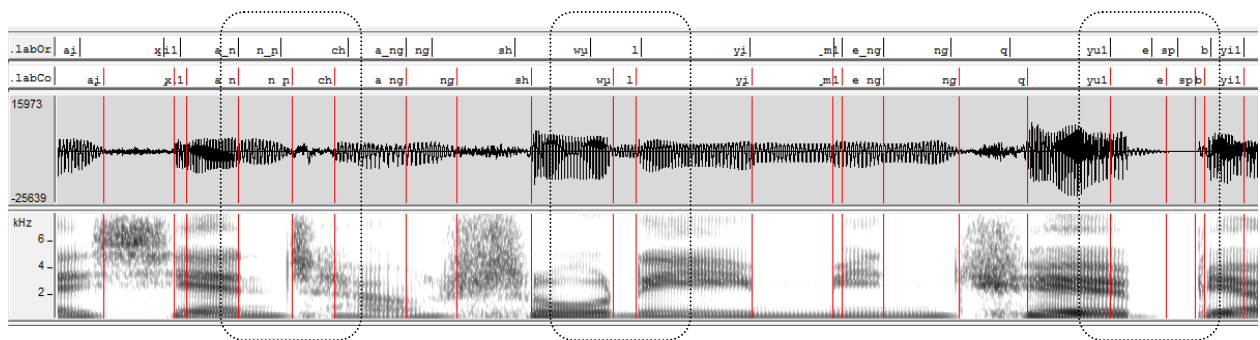


圖 3.12: 自動調整國語語句端點位置實驗結果之範例二，最上方兩列標音位置分別表示 HMM


自動語音分段及修正後之語音標記位置、語音波形、聲譜圖

# 第四章音素端點偵測器架構

本論文展示一個新穎的取樣點式聲學參數建構之音素端點偵測系統，利用本研究所提出之取樣點式的聲學參數描述語音信號的變化特性，並加入英文 TIMIT 語料庫其音素層級之文字轉寫的人為時間標記做為偵測器模型初始化訓練之目標，以半監督式 (semi-supervised) 的方法，訓練音素端點偵測器的模型。4.1 節將說明本論文提出音素端點偵測器架構之概念及系統的建構；4.2 節將會介紹在本研究中所使用的取樣點式聲學參數之抽取方法；4.3 節將介紹音素端點偵測器模型訓練及反覆疊代更新之演算法。

## 4.1 音素端點偵測器架構之設計

### 4.1.1 音素端點偵測系統



儘管在不同語言之中，人類的發音系統之構造對語音的影響，在一段語句內即顯現出其音素的語音特性皆與發音部位以及發音方法有非常大的關聯性。由第三章所述，本論文提出取樣點式聲學參數的聲學特性來描述這些語音信號中不同語音屬性的變化，藉由量測這些變化來找出可能為音素端點的位置，這意謂著進行語音的標記中並不需要完整的音素辨認流程，也不需使用到非常準確的音素標記位置，即可簡化語料庫繁複處理的過程。

端點偵測器以音素層級之人為時間標記文字轉寫來訂定目標函數的兩種轉移狀態，分別為音素端點 (T)、非音素端點 (nT)，對所有由預選擇候選端點 (Candidate Pre-selection) 對應文字轉寫標記目標函數的種類，並用於端點偵測器的訓練。其中，對於每個候選端點其包含了自身端點的聲學特性及其與前後相鄰候選端點之間的音段聲學特性，最後經由多層感知器的學習特性，反覆疊代訓練將音素端點與非音素端點的語音特性做分類，並藉此模型達到音素端點偵測的目的。

本論文所建構之音素端點系統是利用英文 TIMIT 語料庫所提供之人為時間標記的文字轉寫作為音素端點偵測器模型初始化訓練之目標。採用半監督式的訓練方式，來獲得一個端

點偵測器模型。利用訓練後的音素端點偵測器模型，對不同語料庫進行音素端點的偵測，實驗結果將於下章節做分析。圖 4.1 為訓練音素端點偵測系統之流程圖，分為抽取聲學參數以及音素端點模型之訓練方式兩個部分，此兩部分將於 4.2 節、4.3 節作介紹。

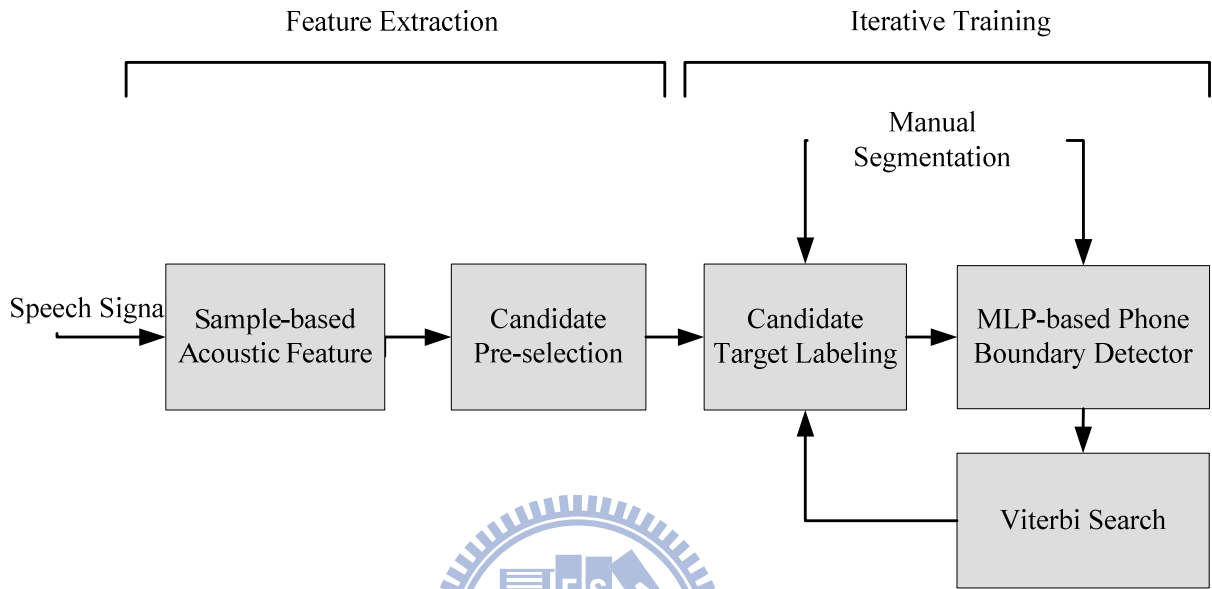


圖 4.1：使用多層感知器架構之音素端點偵測器

## 4.1.2 自動語音分段系統

本論文所建構之自動語音分段系統是分為兩階段式 (Two-stage) 的端點位置修正 (boundary refinement)。第一階段以 MFCC 聲學參數利用 HMM 模型進行強迫對齊而得到初始的語音分段位置；第二階段由本論文提出之取樣點式聲學參數經多層感知器對不同語音單元分類訓練端點偵測器，並依此架構對第一階段所得到之初始語音分段位置做更細部的調整，最後系統輸出對應於語音單元之文字轉寫的自動語音分段位置。圖 4.2 展示了自動語音分段系統之流程圖，其主要與音素端點偵測器架構的差別是在於目標函數的定義。自動語音分段系統之模型描述了語言之音節結構對應至語音分段之關聯性。

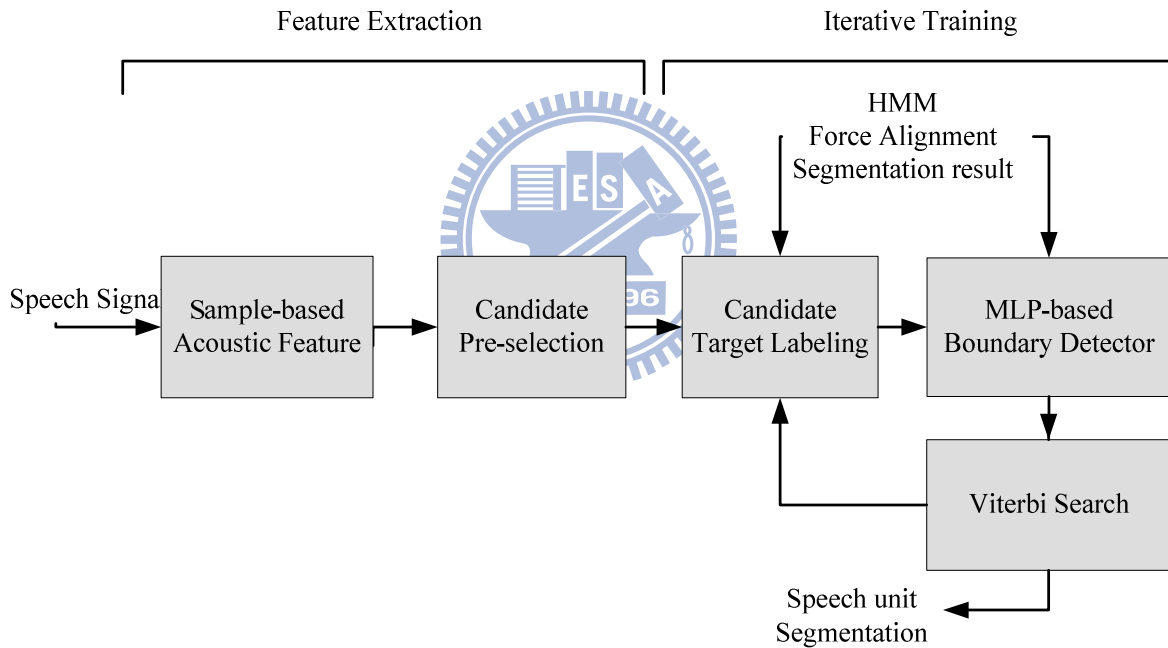


圖 4.2：使用多層感知器架構之自動語音分段系統流程圖

然而，需要做自動語音分段處理的文字轉寫必須根據基本語音單元並依照音節結構來訂定目標函數的種類，以提供端點偵測器的學習。藉由任務的不同來選擇適當的語音單元來進行處理，例如語音合成系統所需要的單元約在聲母/韻母甚至是音節的語音單元；語音辨識系統則可能需要小至音素等語音單元。2.2 節的圖 2.2 顯示出同屬聲調語言之國、客語的音節



結構，在本論文中選擇處理的語音單元為客語語料庫為聲/韻母的語音單元，國語語料庫為類音素以及音節單元。以下將說明選擇不同基本語音之單元其目標函數之訂定方式：

➤ 音節層級

以音節結構之音節層級來訂定語音信號所對應的兩個類別 (class)，分別為靜音 (S) 與音節 (V)，依照不同類別彼此之間的轉移狀態，定義五種目標函數分別是 IS、SV、IV、VS、VV 等轉移狀態，如圖 4.3 表示。每個由抽取聲學參數過程中所得到的候選端點皆須要進行目標函數的標記，圖中之 IS 轉移狀態代表該候選端點仍為靜音狀態，SV 轉移狀態表示該候選端點是由靜音狀態轉換至音節狀態，依此類推...。其中需要特別注意的是圖中 VV 的轉移狀態為表示略過靜音至下一個音節的音節端點。聲調語言中每個音節與音節之間靜音的存在可有可無，為描述此種情形本論文加入 VV 轉移狀態來模擬音節之間無靜音的現象。

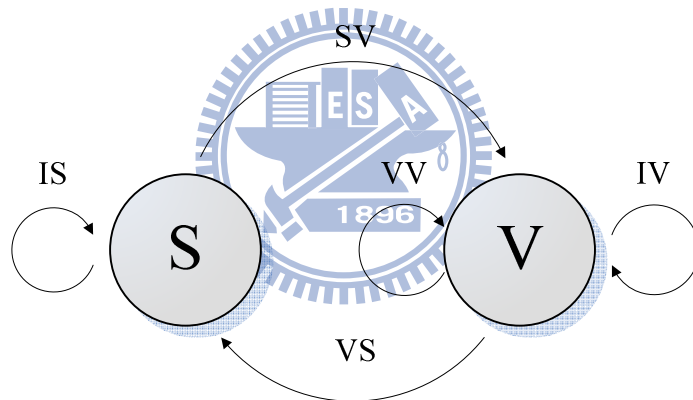


圖 4.3：音節層級目標函數之轉移狀態圖

➤ 聲/韻母層級

以音節結構之聲/韻母層級來訂定語音信號所對應的三個類別，分別為靜音 (S)、聲母 (C) 和韻母 (V)，依照不同類別彼此之間的轉移狀態，定義七種目標函數分別是 IS、SC、IC、CV、IV、VS、VC 等轉移狀態，如圖 4.4 表示。圖中之 IS 轉移狀態代表該候選端點仍為靜音狀態，SC 轉移狀態表示該候選端點是由靜音狀態轉換至聲母狀態，同樣地依此類推...。另外，圖中 VC 的轉移狀態為模擬音節之間無靜音的現象，其代表由韻母與下一個聲母轉移狀態的端點。



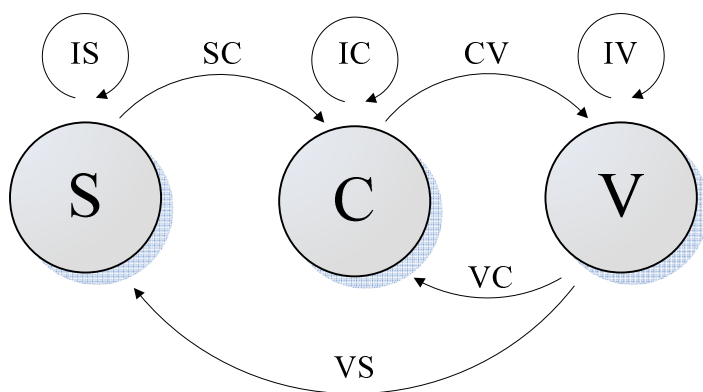


圖 4.4：聲/韻母層級目標函數之轉移狀態圖

➤ 類音素層級

以音節結構之類音素層級來訂定語音信號所對應的四個類別，分別為靜音 (S)、聲母 (C)、韻母 (V) 與鼻音韻尾 (N)，依照不同類別彼此之間的轉移狀態，定義九種目標函數分別是 IS、SC、IC、CV、IV、VN、IN、VS、VC 等轉移狀態，如圖 4.5 表示。另外，圖中為簡化目標函數之個數，本論文將鼻音韻尾至靜音與韻母至靜音的轉移狀態定義為相同的目標函數 (VS)；另外，模擬音節之間無靜音的現象中，本論文亦將鼻音韻尾至聲母與韻母至聲母的轉移狀態定義為相同的目標函數 (VC)。

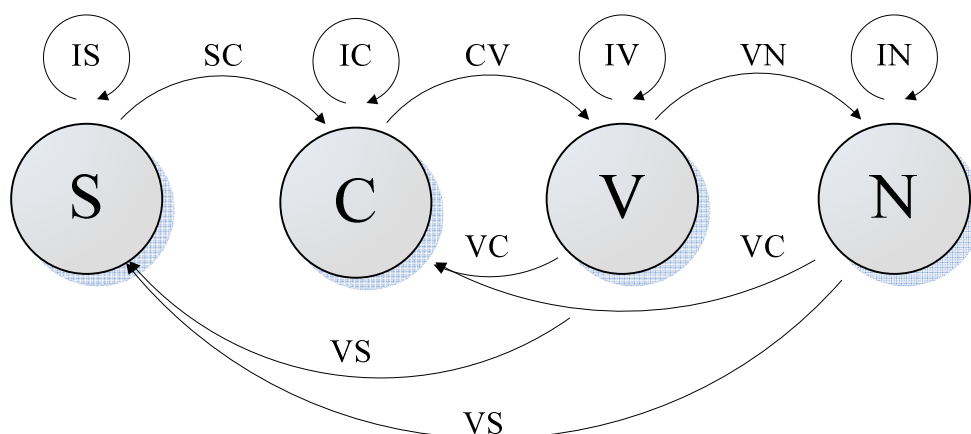


圖 4.5：類音素層級目標函數之轉移狀態圖

由上述不同層級之目標函數轉移狀態的訂定，訓練不同文字轉寫之基本語音單元使用的音素端點偵測器來達到自動語音分段的目的。

## 4.2 聲學參數之萃取

抽取聲學參數之目的是為呈現一段語音信號其特性之表示方式，語音信號內不同之音素有不一樣之特徵而相似音素其特徵也相差不遠。實際上，雖然音素的語音特性會根據不同語者及語者所朗讀文句之內容不一樣進而產生數值上的變化、差異或者是音位變體(Allophone)的效應。但總體來說，其語音屬性卻是不變的，藉由抽取聲學參數的過程，將語音信號中所隱含的聲學資訊提取出來，並依此進行資訊的分析與歸類 (Classification) 以提供進一步的處理。聲學參數的選擇與萃取方式為影響語音辨識效能之重要的前處理步驟，以下將介紹傳統語音研究最常使用的聲學參數與本論文中抽取取樣點式聲學參數的方式。

### 4.2.1 傳統語音聲學參數萃取方式

#### ➤ 線性預測係數 (Linear Predictive Coefficient, LPC)

線性預測通常可視為一個自回歸分析 (Autoregressive analysis)，在許多應用層面如語音編碼、語音合成或是語音辨識等為一個重要的方法。假若將人類之發聲系統來一個建立線性預測的模型，且其為整合發聲器官、口腔形狀和嘴唇發聲之輻射效應的一個全極點模型如下示表示：

$$H(z) = \frac{G}{1 - \sum_{i=1}^p \hat{a}(i)z^{-i}} \quad (4-1)$$

其中  $H(z)$  為系統轉移函數，其系統增益以  $G$  表示並且  $\hat{a}(i)$  為  $p$  階多項式的係數。此模型之輸入信號為一激發序列並以聲帶震動與否來決定是否要加入增益。

對於聲帶震動的激發序列，亦即元音信號，其口腔形狀可視為一個長且細的聲響管 (Acoustic tube) 且其系統轉移函數可用一個全極點的模型來描述。而聲響管之共振頻率可視為語音信號之共振峰，其對應於線性預測全極點模型之極點在頻域之位置，這也就是此模型可描述語音信號頻域上的波封之原因。

然而線性預測的目標即為使預測信號之誤差達到最小，意即找到一組係數 $\hat{a}(i)$ 使得預測信號 $\tilde{x}(n)$ 與原語音信號 $x(n)$ 之均方差值（Mean Squared Error, MSE）最小，此組係數即為線性預測係數。係數求解的方式有很多種，如自相關法（Auto-correlation method）、協方差法（Covariance method）、格型法（Lattice method）等等。因為線性預測係數能有效率地且快速的計算，使得此一聲學參數受到廣泛地使用。

### ➤ 梅爾倒頻譜係數（Mel-Frequency Cepstral Coefficients, MFCCs）

梅爾倒頻譜係數的求取，為將語音信號以少量的數值來模仿耳朵內之基底膜（Basement membrane）其聲音對臨界頻帶（Critical band）的刺激反應，此係數亦表現了人類聽覺系統對音頻是以對數級的感受程度。梅爾倒頻譜係數雖然是一個經過對應聽覺感知的聲學參數，但是其仍是容易且快速計算的參數。

計算梅爾倒頻譜係數步驟如下：

- (1) 將語音信號以視窗函數（Window function）音框化（Frame blocking），通常使用的視窗函數為漢明窗（Hamming window）且音框平移為 5 至 10 毫秒以及音框長度約 10 至 20 毫秒。
- (2) 以快速傅立葉轉換（Fast Fourier Transform, FFT）將音框化後之語音信號轉換至頻域上以得到此音框信號之頻譜。
- (3) 所得到之頻譜能量對應於梅爾刻度頻率曲線之三角帶通濾波器（Mel-scale filter bank）並求得每個濾波器輸出之對數能量。而梅爾刻度頻率曲線與一般頻率的關係式如下：

$$Mel(f) = 2595 \times \log_{10} \left( 1 + \frac{f}{700} \right) \quad (4-2)$$

- (4) 使用離散餘弦轉換（Discrete Cosine Transform, DCT）將對數能量轉換倒頻譜域以求得  $p$  階之梅爾倒頻譜係數。其離散餘弦轉換公式如以下：

$$X_k = \sum_{n=0}^{N-1} x_n \cos \left[ \frac{\pi}{N} \left( n + \frac{1}{2} \right) k \right] \quad k = 0, \dots, N-1 \quad (4-3)$$

自動語音辨識的研究通常使用 10 至 12 個階數之係數即可，其被認為已足夠代表語音信號的特性。再者，由於人嘴唇所發出的聲音受到傳播時輻射效應的影響，使得所聽到的語音之頻譜具有隨著頻率增加而強度降低的特性，頻譜的波封經過頻域與對數的轉換後，造成係數的階數越高而遞減的現象。

## 4.2.1 取樣點式語音聲學參數萃取方式

取樣點式的音素端點偵測架構中，首先使用計算同第 3 章節所述之取樣點式聲學參數，來得到 6 個子頻段信號波封，值得注意的是在此一計算過程當中做了一些適當的調整。即計算時將這 6 個子頻段信號波封輸出加上一個臨限值，此臨限值是為降低每個頻段微弱信號部分的變動影響，例如雜訊。

$$E_i[n] = \begin{cases} \frac{e_i[n]}{\sum_{j=1}^6 e_j[n]}, & \frac{e_i[n]}{\sum_{j=1}^6 e_j[n]} > \eta \\ \eta, & \text{otherwise} \end{cases} \quad (4-4)$$



從語音信號中抽取聲學參數之後，為了減少在端點偵測器內過於龐大的資料計算量，經由預選擇即如同 3.1.3 節所敘述，藉由簡單設定一個臨限值 ( $Th_d$ ) 的方法來挑選可能較大之音素端點位置；由於頻譜 KL 距離在挑選出語音信號相鄰時間中的變化上是一種很好的量測方式，故若頻譜 KL 距離滿足下式：

$$d_{KL}[n-1] < d_{KL}[n], d_{KL}[n] > d_{KL}[n+1] \text{ and } d_{KL}[n] \geq Th_d \quad (4-5)$$

則代表為挑選出來的候選端點值，最後得到這一序列音素的候選端點， $\{c_j; j=1, \dots, N\}$ 。

經過預選擇步驟後，在此實驗過程中依照觀察頻譜 KL 距離與人為時間標記之間的關係發現一些現象，舉例來說對於人為時間標記中之摩擦音至母音、流音之間的音素轉換端點，在聲譜圖中可觀察到端點兩邊頻譜信號分佈的差異極大如圖 4.6 中的 (k-l)、(t-ix) 之轉換端點，圖中可以看到人為時間標記的位置並不一定是相鄰區域中頻譜 KL 距離局部極大值的端

點，而是黑色箭頭所指向的端點；另外，圖中偏右旁的 (k-l) 音素轉換端點之相鄰區域中並無特別大的頻譜 KL 距離，那麼要如何選擇最適當的音素候選端點能減少訓練音素端點偵測器所需要達到收斂的次數？此問題即為先前所描述其人為時間標記之語料庫其標音員之主觀性所產生時間標記位置之不一致性的問題。

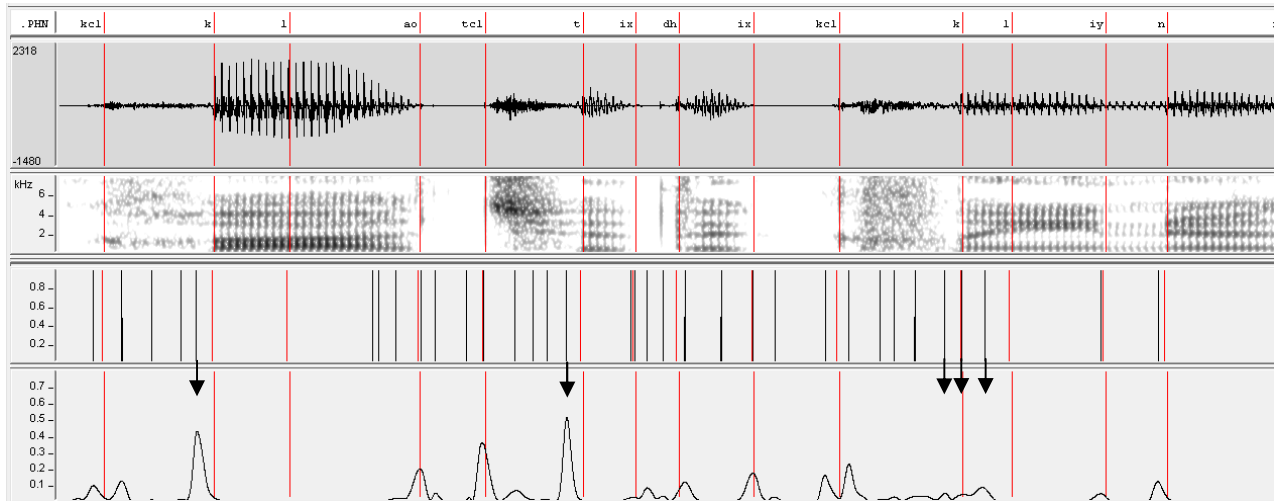


圖 4.6：調整音素候選端點之範例，由上至下分別表示音素層級之人為時間標記的文字轉寫、語音信號、聲譜圖、音素候選端點、頻譜 KL 距離

因此，本論文提出一個演算法用以挑選出候選端點序列中最佳的音素候選端點作為半監督式學習的目標 (Target)。

其演算法的敘述如下：

- (1) 在人為之時間標記音素端點之相鄰區域選擇適當的範圍，本論文使用相鄰音素端點之中點作為上限 (Upper bound, UB) 與下限 (Lower bound, LB) 且前後以不超過 30 毫秒的範圍作為挑選候選端點的區域。
- (2) 在區域  $R$  內頻譜 KL 距離挑選出來之候選端點即為第  $i$  個音素端點之候選端點子序列  $\{c_{i,j}; j=1, \dots, k\}$ ，並將此子序列依候選端點與該音素端點之距離由近至遠排序。
- (3) 將排序好的候選端點子序列依照臨限值<sup>3</sup> ( $Th_c$ ) 判斷，得到此序列中最佳的音素候

<sup>3</sup>經由觀察頻譜 KL 距離對應語音信號變化之數值我們設定一臨限值，假若其候選端點之頻譜 KL 距離大於臨限值我們便認為其端點是極有可能為音素端點的位置。

選端點  $c_{i,j}$ ，並標記此候選端點為第  $i$  個音素端點所要學習的目標。

- (4) 重複(1)、(2)、(3)的步驟直至所有音素端點皆經過計算後，求得所有最佳之音素候選端點並完成學習目標的標記。

藉由候選端點會將語音信號分割成很多音段 (Segment)，反而言之，這些音段相較於由頻譜 KL 挑選之音素候選端點的語音特性是可視為穩定的，故即可使用這些音段之語音信號求取一些音段式 (Segment-based) 的聲學參數來描述候選端點兩旁之語音特性，以協助進行音素之端點偵測。

首先，本論文使用音段式的子頻段信號波封 (Segmental sub-band signal envelope) 來表示 2 個相鄰的音段  $[c_{k-1}, c_k]$ 、 $[c_k, c_{k+1}]$  內其語音信號在頻譜的分佈情形，在此以下圖 4.7 來作說明。圖中候選端點  $k$  之高度表示頻譜 KL 距離數值之大小，其前、後音段 (Segment k-1、Segment k) 則分別表示在候選端點間其語音特性的狀態，假若候選端點相鄰兩旁音段之頻譜信號分佈差異極大，代表其語音信號轉變而造成其分佈差異，那麼即可增加此一輔助資訊來提升音素端點偵測之效能。因此，本研究定義候選端點相鄰音段  $ES_i(k)$  為在第  $k$  個音段  $[c_{k-1}, c_k]$  中其子頻段信號波封經正規化後的平均值，如下式：

$$ES_i[c_{k-1}, c_k] = \left( \sum_{n=c_{k-1}+\Delta}^{c_k-\Delta} E_i[n] \right) / (c_k - c_{k-1} - 2\delta) \quad (4-6)$$

其中  $\delta$  表示與候選端點  $k$  相距的取樣點個數。

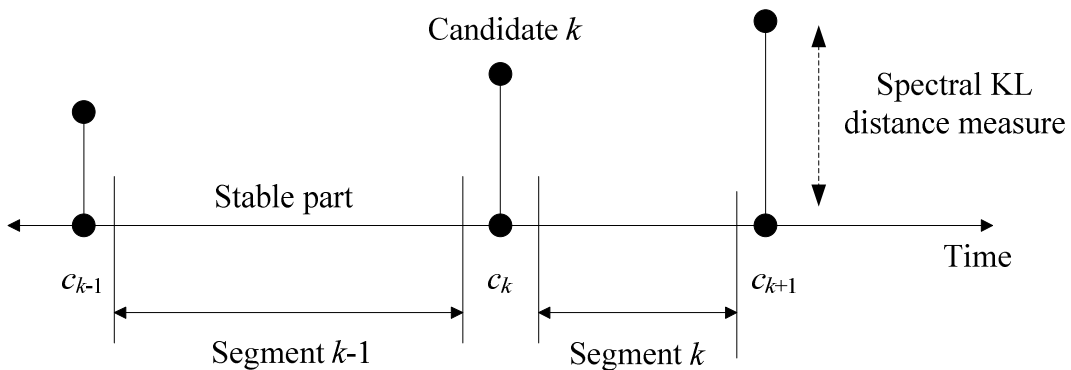


圖 4.7：利用候選端點將語音信號分割成片段的示意圖



接著，考慮相鄰候選端點之時間關聯性與其端點間語音特性之相關性，對於每個候選端點建立一個 38 維的聲學參數向量，對於第  $k$  個候選端點， $c_k$ ，其聲學參數向量包括以下聲學參數：

(1) 目前候選端點及前、後候選端點之參數：

$$\left( E_i[c_k], \Delta E_i[c_k]; i=1, \dots, 6 \right), \Delta E_0[c_k], \left( d_{KL}[c_j], H_s[c_j], \Delta H_s[c_j] \right); j=k-1, k, k+1$$

其中  $\Delta E_i[c_k]$ 、 $\Delta E_0[c_k]$  與  $\Delta H_s[c_j]$  分別為第  $i$  個經正規化之子頻段信號波封、語音信號波封與頻譜熵之一階差量。

(2) 目前音段及前、後音段之參數：

$$\left( ES_i[c_{k-1}, c_k], ES_i[c_k, c_{k+1}]; i=1, \dots, 6 \right), c_k - c_{k-1}, c_{k+1} - c_k$$

其中  $c_k - c_{k-1}$ 、 $c_{k+1} - c_k$  表示目前端點與前後相鄰端點之時間資訊。

(3) 使用 2 個指標指出此候選端點是否為此候選端點序列之第一個或者最後一個端點。

最後，由語音信號所抽取之每個聲學參數向量皆存在聲學參數檔案內，以提供後級音素端點偵測器之訓練使用。圖 4.8 展示了抽取聲學參數演算法的整體架構。

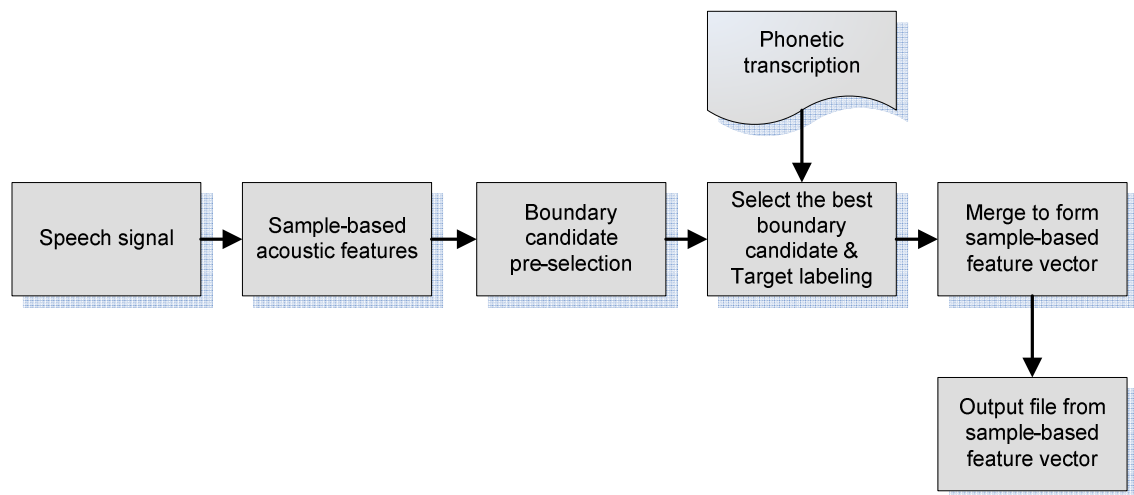


圖 4.8：聲學參數抽取演算法的系統架構圖



## 4.3 模型訓練

完成語音之聲學參數萃取後，本節中將介紹音素端點偵測器模型之演算法，本研究使用 NIKO toolkit[13] 多層感知器之類神經網路架構，將基於使誤差值最小化的準則（Error minimization）採用倒傳遞演算法（Back-propagation algorithm）將先前所建置之取樣式聲學參數進行參數資料的分群訓練與模型目標函數的更新。以下分別會在 4.3.1 介紹初始化（initialization）訓練以類神經網路建構的音素端點偵測器模型；4.3.2 節敘述反覆疊代（iteration）更新模型參數之演算法。

### 4.3.1 多層感知器之類神經網路架構

類神經網路是一種基於模擬人類腦中之神經元相互傳遞資訊情況而發展的處理技術，其可利用輸入與輸出所組成的資料對（training pair）來建立系統模型，並經系統自我學習、推估甚至是做決策。因此，一個類神經網路其包含了許多人工神經元（neuron）與神經元彼此間之鏈結（links），依此組成各種不同的神經網路。而依照不同的鏈結架構可分為前饋式（feed-forward）與回饋式（feed-backward）的類神經網路，本論文使用屬前饋式類神經網路的多層感知器，將在以下作其運作之介紹。

#### 4.3.1.1 神經元之結構

神經元（如圖 4.9）是類神經網路中最基本的單位，其負責處理資料之間的關係。而在網路中的某個神經元，會收到一個至數個不等的輸入參數，神經元對於每個輸入參數  $x_j$  經處理後，會依照輸入參數的重要程度而加乘上一個權值（Weight,  $w$ ）。最後，神經元將所有經加乘權值過之所有輸入參數累加起來，接著與神經元中的偏移量（Bias,  $b$ ）相加為一淨值（*net*），經由激發函數（activation function）之特性將淨值轉換成模擬人類神經元的訊號。其神經元的數值表示公式如(4-8)所示：

$$net = \sum_j w_j x_j + b \quad (4-7)$$

$$a = f\left(\sum_j w_j x_j + b\right) \quad (4-8)$$

其中  $a$  表示輸入變數經神經元處理後由激發函數所產生的數值。

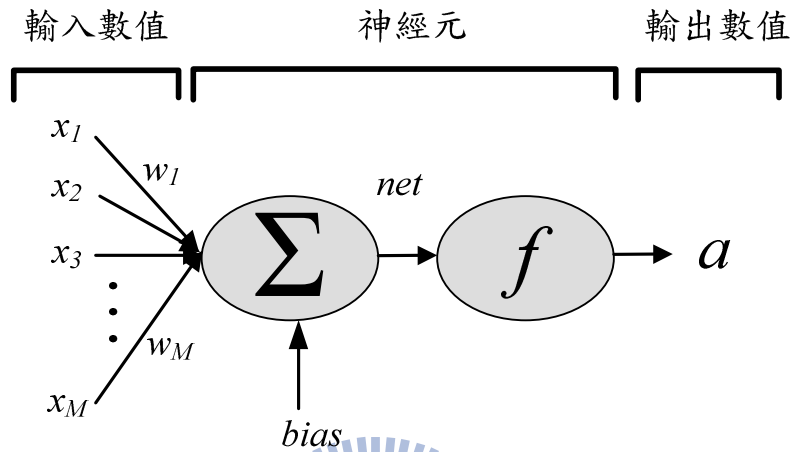


圖 4.9：神經元輸入輸出關係圖

神經元之激發函數可依照面臨的問題不同來選擇其需要之特性，主要可分為線性函數或非線性函數兩種。而激發函數的種類有很多，本論文使用雙曲正切函數（Hyperbolic Tangent function）作為神經元之激發函數，其數學式表示如下：

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (4-9)$$

圖 4.10 所示，神經元之激發函數輸出之數值範圍在-1 至+1 之間。

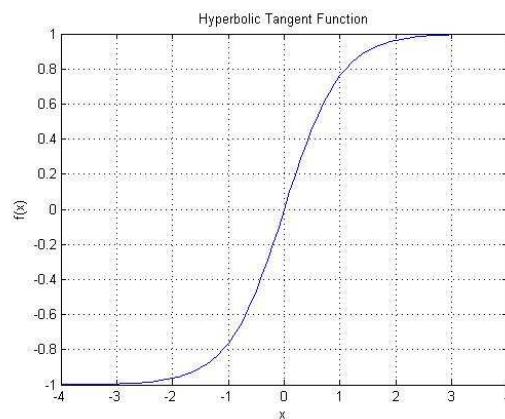


圖 4.10：雙曲正切函數之激發函數曲線圖

### 4.3.1.2 網路之結構

前饋式類神經網路結構如圖 4.11 所示，數個平行輸入參數的神經元並聯組成為網路的結構層 (Layer)，再經由數個層之間的神經元相互鏈結進而串聯成一個前饋式的類神經網路。依照各層不同的特性可將其分為輸入層、隱藏層與輸出層三種結構層的分類。

- (1) 輸入層：輸入層之每個神經元以一個輸入參數數作為輸入值，以提供後級下一層之每個神經元的輸入以進行訓練。意即此層之神經元作用類似暫存器的效用，不具有運算的功能。輸入層神經元數目即為輸入參數向量的個數，本論文所使用是參數向量為 38 維，故輸入層之個數為 38 個。
- (2) 隱藏層：隱藏層介於輸入層與輸出層之間，而隱藏層神經元的個數並沒有一定的限制，隱藏層的層數與神經元的個數是依照資料的複雜程度所測試後決定。本論文所使用之隱藏層的層數為一層，隱藏層神經元的個數為 75 個。
- (3) 輸出層：每個神經元的輸出值為類神經網路經過訓練所產生的目標函數輸出值，輸出層神經元的數目應等同於網路的輸出值總數，其依照端點偵測任務 (task) 不同而調整不同目標函數個數。

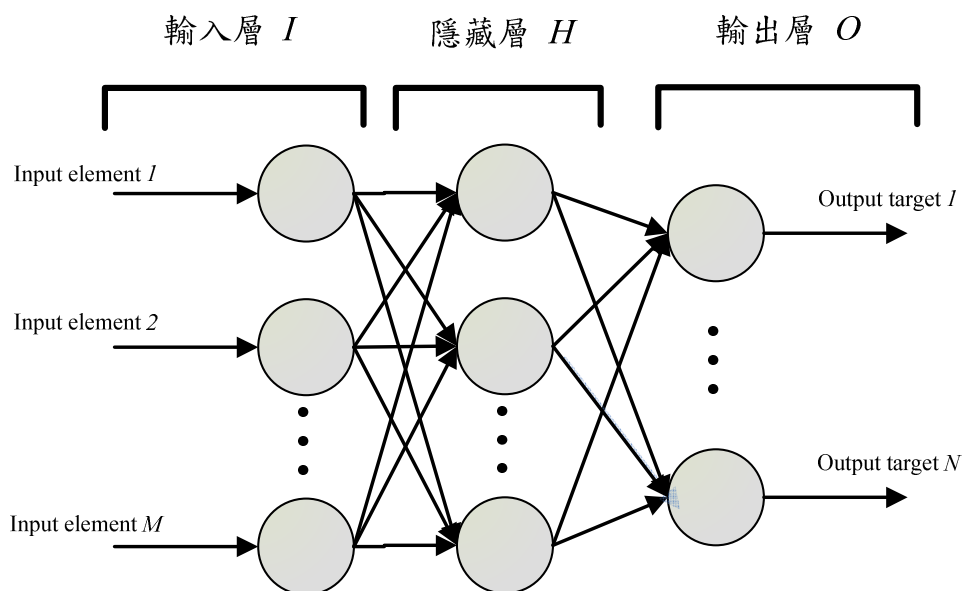


圖 4.11：多層前饋式類神經網路結構範例

### 4.3.1.3 網路學習方法

本論文以半監督的訓練學習方法，藉由批次訓練<sup>4</sup> (batch learning) 過程中利用倒傳遞演算法來調整各神經元之間鏈結的強弱，亦即透過更新加權值來描述各神經元間之關連性，故每個輸入參數向量皆需要有對應的目標函數以提供網路的學習，並在一定次數的批次訓練後再更新目標函數繼續進行訓練。

在此，本論文定義每一層之神經元個數分別為  $N_1$ 、 $N_2$ 、 $N_3$ ； $w_{ij}^{(\alpha\beta)}$  表示為第  $\beta$  層內之第  $j$  個神經元至第  $\alpha$  層內之第  $i$  個神經元的加乘權值； $I_i(n)$ 、 $H_i(n)$ 、 $O_i(n)$  分別表示輸入之第  $n$  筆輸入參數向量其輸入層、隱藏層及輸出層的第  $i$  個神經元輸出之數值； $b_i^{(k)}$  為第  $k$  層之第  $i$  個神經元的偏壓值；激發函數  $f$  由先前所述為雙曲正切函數，則隱藏層經加權處理後之輸出為：

$$H_j(n) = f\left(\sum_{l=0}^{N_1-1} w_{jl}^{(21)} I_l(n) + b_j^{(2)}\right) \quad (4-10)$$

其中  $i=0, \dots, N_3-1$ 。多層感知器之輸出為：

$$O_i(n) = f\left(\sum_{j=0}^{N_2-1} w_{ij}^{(32)} H_j(n) + b_i^{(3)}\right) \quad (4-11)$$

其中  $j=0, \dots, N_2-1$ 。

而本論文的目的是為利用多層感知器之音素端點偵測模型來辨別輸入參數進而達到最佳的效能。因此訓練過程必須將類神經網路的輸出與音素端點偵測器設定目標函數之誤差值最小化。定義誤差函數 (Error function) 如以下：

$$E(n) = \frac{1}{2} \sum_{i \in N_3} (T_i(n) - O_i(n))^2 = \frac{1}{2} \sum_{i \in N_3} e_i^2(n) \quad (4-12)$$

<sup>4</sup> 類神經網路學習過程中其更新加權值的方式可分為兩種，批次訓練為經過全部的輸入樣本才更新權值，而循序訓練(sequential training)則為在處理完一個輸入樣本便立即更新加權值的方式。

$$E = \sum_n E(n) \quad (4-13)$$

其中， $E(n)$  為輸入第  $n$  筆之參數向量其類神經網路的誤差， $T_i(n)$  為輸出層之第  $i$  個神經元目標函數， $N_3$  為輸出節點數， $e_i(n)$  為輸出層之第  $i$  個神經元目標函數與輸出的差值， $E$  為所有輸入參數向量輸出誤差之總合。

倒傳遞演算法之基本原理為利用最陡坡降法 (Steepest decent method)，藉由適當地調整神經元之間的，以達到將誤差函數最小化的目的。如以下表示：

$$E(W + \Delta W) < E(W) \quad (4-14)$$

為得到一適當的權值差量  $\Delta W$ ，對誤差函數對權值作偏微分以求得相對的梯度。則誤差函數對輸出層至隱藏層的偏微分所得之權值差量  $\Delta w_{ij}^{(32)}$  如(4-15)：

$$\Delta w_{ij}^{(32)} = -\eta \frac{\partial E}{\partial w_{ij}^{(32)}} = \eta \cdot \delta_i \cdot H_j(n) \quad (4-15)$$

其中  $\delta_i = e_i(n) \cdot f'(net_i^3(n))$ ， $net_i^3(n) = \sum_{j=0}^{N_3-1} w_{ij}^{(32)} H_j(n) + b_i^3$  為第三層之第  $i$  個神經元之淨值。 $\eta$  為學習速率 (Learning rate)。相同地，我們也可由隱藏層至輸入層得到誤差函數對加權值偏微分所得之權值差量  $\Delta w_{jl}^{(21)}$  表示為：

$$\Delta w_{jl}^{(21)} = -\eta \frac{\partial E}{\partial w_{jl}^{(21)}} = \eta \cdot \delta_j \cdot I_l(n) \quad (4-16)$$

其中  $\delta_j = e_j(n) \cdot f'(net_j^{(2)}(n))$ ， $net_j^{(2)}(n) = \sum_{l=0}^{N_1-1} w_{jl}^{(21)} I_l(n) + b_j^{(2)}$  為第二層之第  $j$  個神經元之淨值。

因此在得到層與層之間加權值調整過後的權值差量，為避免權值差量的變化不穩定，加入衝量係數  $\mu$  與前一次的權值差量  $w_{ij}^{(\alpha\beta)}(t-1)$  帶入更新公式：

$$\Delta w_{ij}^{(32)}(t) = \mu \cdot \Delta w_{ij}^{(32)}(t-1) - \eta \frac{1}{N} \sum_{n=0}^{N-1} \Delta w_{ij}^{(32)} \quad (4-17)$$

$$\Delta w_{jl}^{(21)}(t) = \mu \cdot \Delta w_{jl}^{(21)}(t-1) - \eta \frac{1}{N} \sum_{n=0}^{N-1} \Delta w_{jl}^{(21)} \quad (4-18)$$

其中  $N$  為輸入參數向量之總數。

最後，每經過一次批次訓練則更新加權值，其公式為

$$w_{ij}^{(32)}(t) = w_{ij}^{(32)}(t-1) + \Delta w_{ij}^{(32)}(t-1) \quad (4-19)$$

$$w_{jl}^{(21)}(t) = w_{jl}^{(21)}(t-1) + \Delta w_{jl}^{(21)}(t-1) \quad (4-20)$$

學習速率的選擇是非常重要的，其影響著最陡坡降法修正加權值的幅度，假若學習速率太小會造成收斂次數的提高而拉長時間；學習速率過大則會在收斂的過程中產生震盪，即誤差函數無法遞減而收斂。經過實驗的測試，本論文所使用類神經網路之參數初始值如下表：

表 4.1：類神經網路參數初始設定值

偏壓值	學習速率	衡量
0.01	0.00001	0.9

### 4.3.2 反覆疊代

在有 TIMIT 語料庫人為時間標記之文字轉寫作為模型初始化訓練後，為實現半監督式的訓練方式，以下將介紹訓練音素端點偵測器模型反覆疊代的步驟，其流程圖如圖 4.12：

➤ **Step1：將多層感知器輸出之概似度 (likelihood) 正規化為機率**

依照目標函數的個數將多層感知器之輸出層對應每個輸入聲學參數向量所產生之概似度作正規化，則得到該參數向量在各個目標函數機率。

➤ **Step2：更新文字轉寫之自動時間標記**

接著，使用維特比搜尋演算法 (Viterbi search algorithm) 重新將文字轉寫作強迫對齊，以得到一個更新後的自動語音分段位置。

➤ **Step3：重新標記目標函數**

在有一個經重新自動分段後的文字轉寫，由文字轉寫內的時間標記將端點位置再重新標記目標函數，並作為下一次多層感知器之學習目標。

➤ **Step4：更新多層感知器之目標函數**

置換多層感知器的目標函數，繼續訓練音素端點偵測器之模型。

➤ **Step5：重覆 Step1 到 Step4 至收斂**

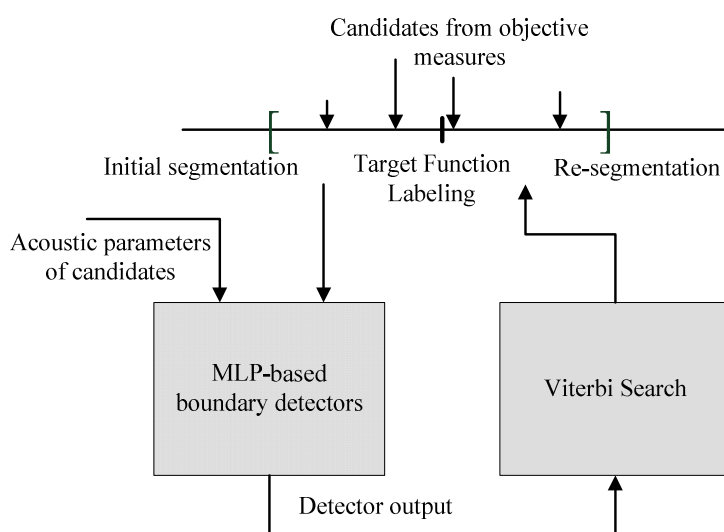


圖 4.12：音素端點偵測器模型反覆疊代之流程圖




# 第五章：實驗結果

在本章節中將應用第三章所使用的取樣點式聲學參數以及第四章所敘述之音素端點偵測器與自動音素分段的訓練演算法於四種不同語料庫（包含英文 TIMIT 語料庫、客語四縣語料庫及國語的 TCC-300 語料庫以及 treebank 語料庫）之實驗觀察結果及效能分析且藉由調整不同子頻段的加權值以改良頻譜 KL 距離，達到最佳的實驗結果。5.1 節將呈現應用於 TIMIT 語料的實驗結果分析；5.2 節將分析國語 TCC-300 及 treebank 語料庫中隊應不同基本語音單位的自動語音分段結果分析；5.3 節將會針對客語語料庫的實驗觀察結果做比較；5.4 改良頻譜 KL 距離以達到更好的偵測效能並作討論。

## 5.1 使用 TIMIT 語料庫之實驗結果

### 5.1.1 音素端點偵測實驗結果分析



使用 TIMIT 語料庫來驗證本論文提出音素端點偵測器的偵測效能，並依照 TIMIT 語料庫所建議訓練語料 4620 個語句及測試語料 1680 個語句的分類，用於音素偵測實驗。首先，表 5.1 統計了訓練語料與測試語料所處理的語音取樣點、音素邊界候選端點（Candidate）以及語料中所要偵測之音素邊界總數（Phone boundary）的數據，由此可推得訓練語料約 1314 個語音取樣點也就是平均約 82.125 毫秒有一個音素端點的存在，而測試語料則是平均每音素端點相隔約 82.83 毫秒，皆與平均音素長度為 50 至 100 毫秒或是約為 5~10 個音框長度的統計量相符；透過 3.1 節適當調整濾波器階數係數並使用 4.2 節萃取出樣點式聲學參數時設定臨限值挑選音素邊界候選端點的方法，分別在訓練語料及測試語料挑選出 509240 與 182858 個可能為音素邊界的候選端點，以提供音素端點偵測器的訓練及實驗結果的統計，最後比對經人為標記的音素層級之文字轉寫而得到偵測音素邊界端點其誤報率與偵測漏失率相等時之錯誤率（Equal error rate, EER）效能為 13% 與 14.5%。而偵測漏失率與誤報率的定義如下式表示：

偵測漏失率為未偵測到之音素邊界端點個數  $D$  在總音素邊界端點個數  $N$  中所佔的比例。

$$\text{Miss Detection rate} = \frac{D}{N} \times 100\% \quad (5-1)$$

誤報率表示誤偵測為音素邊界端點個數  $I$  在總音素邊界端點個數  $N$  與  $I$  之總和中所佔的比例。

$$\text{False Alarm rate} = \frac{I}{I+N} \times 100\% \quad (5-2)$$

表 5.1：TIMIT 語料庫的統計資料結果

TIMIT corpus	Sample	Candidate	Phone boundary	EER
Training part	226727341	509240	172461	13.0%
Test part	82786737	182858	62466	14.5%

在測試語料中所挑選出的候選音素端點，可藉由加上不同的臨限值來控制音素端點偵測器所偵測的音素端點個數，因此實驗中對應不同的臨限值描繪出誤報率與偵測漏失率的對應曲線圖為圖 5.1 所表示，圖中黑色三角點為 Rabiner 在數據中近乎 EER 的數值點，而本論文測試語料與訓練語料的實驗結果分別以藍色實線和紅色虛線表示。然而，誤報率與偵測漏失率為成反比的，在本論文音素端點偵測的觀點中，誤報率的增加代表著有更多音素候選端點被誤認為音素邊界端點的可能性被提高，但音素候選端點是以評量相鄰語音取樣點頻譜差異的頻譜 KL 距離所挑選出來，有些音素的連音現象造成不明顯的頻譜變化，這些部分為較難偵測的音素端點，藉著調降臨限值使誤報率增高，造成對應較難偵測的音素邊界端點也可一併偵測出來，進而減低音素端點偵測的漏失。音素端點偵測的目標為減低人為標記語料庫的繁複過程，過大的偵測漏失率即為音素偵測實驗最不想見的結果。在此，找出誤報率與偵測漏失率之間的取捨平衡點亦即當誤報率與偵測漏失率相同，作為實驗結果的比較方式。

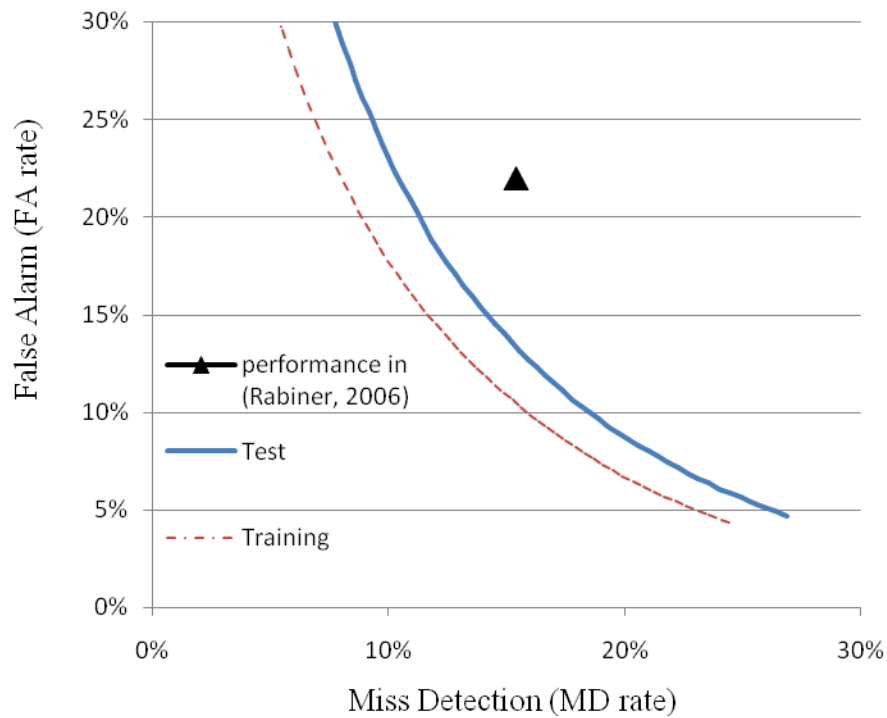


圖 5.1：音素端點偵測器於 TIMIT 語料庫誤報率與偵測漏失率之對應曲線圖

接著，為了能與傳統音框式方法比較實驗的結果，我們統計每 5 毫秒所包含到偵測音素邊界的比例，並計算被偵測到音素端點落在相同或是相鄰音框之內的包含比例，以評量本論文之音素端點偵測器之效能好壞。其中表 5.2 顯示在 EER 的情形下，偵測到的音素邊 endpoint 在不同絕對偏差值內（5、10、15 毫秒）的包含比率，而在相同音框內為 41.72%，相鄰音框範圍內為 87.32%，兩種評量之實驗結果皆優於 Rabiner（27%/ 10ms, 70%/ 20ms），可易見時間解析度較細的取樣點式的音素端點偵測方法有較高的效能。圖 5.2 顯示了音素端點偵測器之實驗結果與人為標記之間的差異在不同絕對偏差值的差異的區間內，佔有總音素端點個數的比例。絕對偏差值越小代表著與人為標記位置越相近，亦表示偵測出之音素候選端點越準確。

表 5.2：使用音框式計算音素邊界偵測結果的方式的統計結果，音框平移為 10ms

5ms	10ms	15ms	In the same frame	In $\pm 1$ frame
43.10%	76.31%	88.37%	41.72%	87.32%

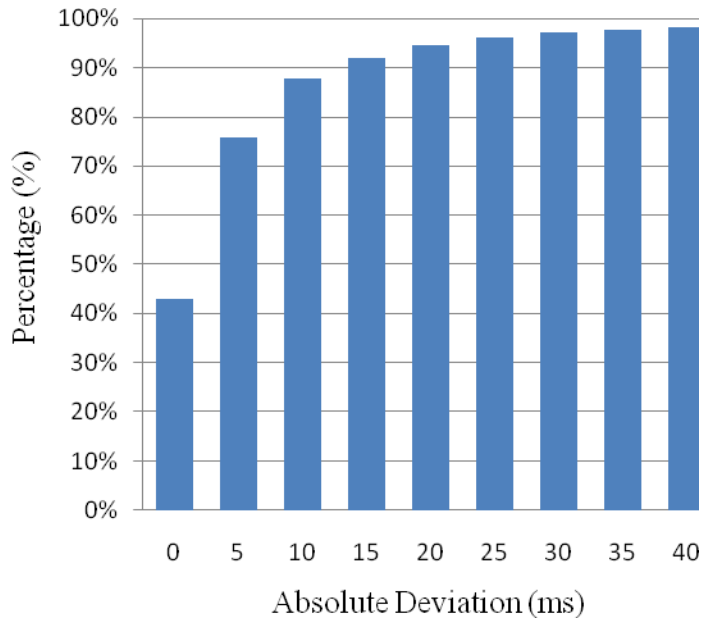


圖 5.2：音素端點偵測器實驗結果與人為標記之絕對偏差值直方圖

由先前所述，有些音素的連音現象其頻譜為平滑的變化，使得這些音素端點非常地難偵測，以下本論文也列舉出觀察語音波形時較難辨別音素轉換對其音素端點偵測的數據。在 EER 的情況下，前後音素均為母音或流音的音素邊界端點有 28.69% 沒有偵測到，而本論文所使用的濾波器組是用於發音方法 (pronunciation manner) 的辨別以及語音中 (landmark) 的偵測。因此，當音素邊界端點的前後音素為相同的發音方法時，偵測漏失率則會高於平均，如前後音素之發音方法為摩擦音的偵測漏失率為 32.32%，而發音方法為鼻音之偵測漏失率為 58.22%。

由以上僅取數種前後音素為相同發音方法的結果，可證實在音素偵測的效能評比上想同發音方法之偵測漏失率確實較平均高上許多，以下針對實驗結果來分析前後音素為相同或是不同發音方法時，其結果對應偵測漏失率和誤報率的變化。在此依照附錄一之分類來統計 TIMIT 訓練及測試語料中的音素對應於不同的發音方法的數量並以表 5.3 左側所示。表 5.3 中的塞擦音的數量相較於其他發音方法少得多，其原因是由於塞擦音包含的音素分類只有 (jh)、(ch) 兩種。另外，表 5.3 的右側為表示測試語料中前後音素相同與不同發音方法的統計資料，很清楚地前後音素無相同塞擦音之發音發音方法。且觀察語音波形時，前後音素為靜音的相同發音方法，如 (h#-tcl) 等，標音員標記音素位置時可能會有較大不一致性的

現象產生；另一方面語音信號更無明顯的變化會使得沒有任何音素候選端點產生，造成偵測漏失率增高。

表 5.3：TIMIT 語料庫中發音方法與前後音素不同發音方法之統計資料

TIMIT corpus	Training	Test	Test corpus	Same	Different
Manners	count	count	Manners	count	count
Stop	25871	9176	Stop	87	9089
Affricates	2031	631	Affricates	0	631
Fricatives	21424	7724	Fricatives	307	7417
Nasals	14157	5104	Nasals	79	5025
Glides	20257	7822	Glides	387	7435
Vowels	57463	20911	Vowels	1497	19414
Silence	35877	12777	Silence	253	10844

#### ➤ 偵測漏失率分析

本論文所提出之方法為利用取樣點式參數的萃取，依照音素變化時語音信號在頻譜之間的變化程度來進行音素邊界端點偵測，若相鄰音素之頻譜變化的程度越大，則越可能被偵測為音素的邊界。由下表 5.4，可以看到相鄰音素是相同與不同的發音方式對照下，實驗結果觀察發現不同的發音方式相較於相同發音方式其大部分之偵測漏失率都有大幅降低的現象。因此以下將針對偵測漏失率較高的摩擦音、鼻音、母音以及靜音等數種發音方法來提出討論。

##### (1) 前後相鄰音素為摩擦音

摩擦音發音時會由於發音器官彼此靠攏而形成狹窄的氣流通道，使得氣流通過通道時造成摩擦產生出聲音，如發出 s 的音必須讓氣流通過閉合牙齒之間的縫隙來產生。摩擦音在頻譜上的分佈多集中在高頻部分。圖 5.3 舉出前後音素為 (k、s) 皆屬於摩擦音的分類，由音素端點偵測器輸出概似度的觀察中，在 (k、s) 音素的區間中所有的音素候選端點之概似度皆非常地低，亦即偵測器不認為這些候選端點是音素的端點。

表 5.4：相鄰音素在相同與不同的發音方法之偵測漏失率

符號\*代表無此組合方式

Test set	MD rate (current and next)	
	same	different
Manners		
Stop	23.0%	16.9%
Affricates	*	6.8%
Fricatives	29.0%	8.0%
Nasals	58.2%	18.4%
Glides	18.3%	22.1%
Vowels	37.7%	11.6%
Silence	66.4%	10.2%

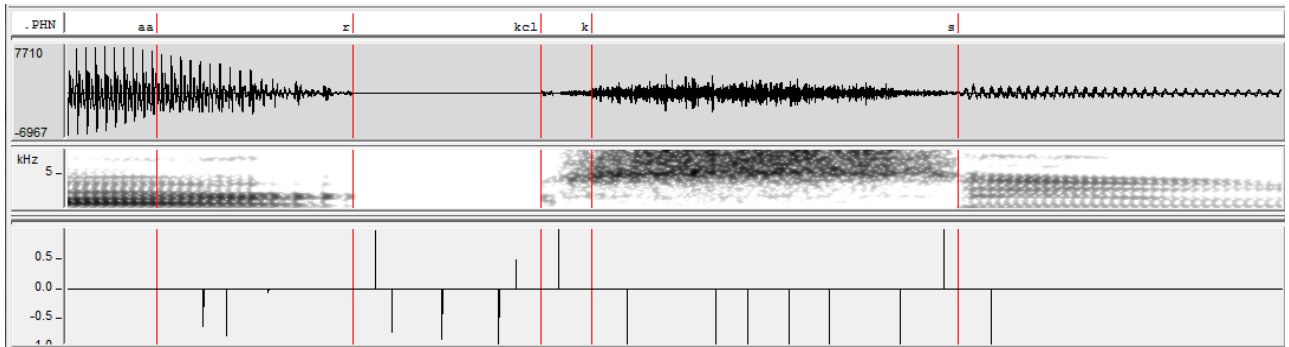


圖 5.3：音素端點偵測前後音素為摩擦音之範例，由上至下分別表示音素層級之人為時間標記的文字轉寫、語音信號、聲譜圖、音素端點偵測器輸出之概似度

## (2) 前後相鄰音素為鼻音

鼻音發音時口腔中的軟顎下垂，造成氣流無法通往口腔而轉往通過鼻腔發音，如發出 m 的音時，須雙唇緊閉來讓氣流通過鼻腔產生，也因為如此使得鼻音在頻譜上的分佈多集中在聲譜圖之低頻部分。圖 5.4 舉出前後音素為 (m、n) 皆屬於鼻音的分類，在 (m、n) 音素的區間中，相鄰音素頻譜間平滑的變化造成音素候選端點的個數較少；僅觀察語音波形也亦難標記正確的音素端點位置，這也就是前後音素為鼻音時偵測漏失率較高的原因之一。即便音素端點偵測器輸出概似度藉由調整臨限值後，增加偵測出候選端點之個數，其音素候選端點仍與人為標記位置有一段誤差存在。



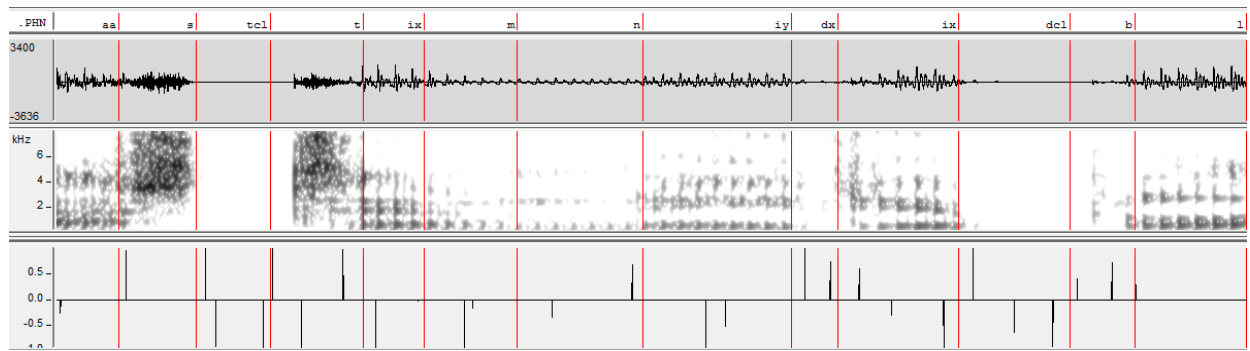


圖 5.4：音素端點偵測前後音素為鼻音之範例，由上至下分別表示音素層級之人為時間標記的文字轉寫、語音信號、聲譜圖、音素端點偵測器輸出之概似度

### (3) 前後相鄰音素為母音

母音是氣流由肺通過聲帶時，使聲帶產生週期性的震動且讓氣流不受阻礙地通過口腔通道，再以舌頭或是雙唇的調整而發出聲音。不同口腔通道、舌頭位置等所發出的母音在頻譜上亦有不同的分佈，但在時域上的語音波形中皆可明顯觀察出週期性的訊號。圖 5.5 舉出前後音素為 (er、axr) 皆屬於母音的分類，相鄰音素頻譜間平滑的變化產生的音素候選端點個數不多，就算偵測器輸出概似度藉由調整臨限值後，增加偵測出候選端點之個數，其音素候選端點仍與人為標記位置有一段誤差存在；同樣觀察語音波形也亦難標記正確的音素端點位置。

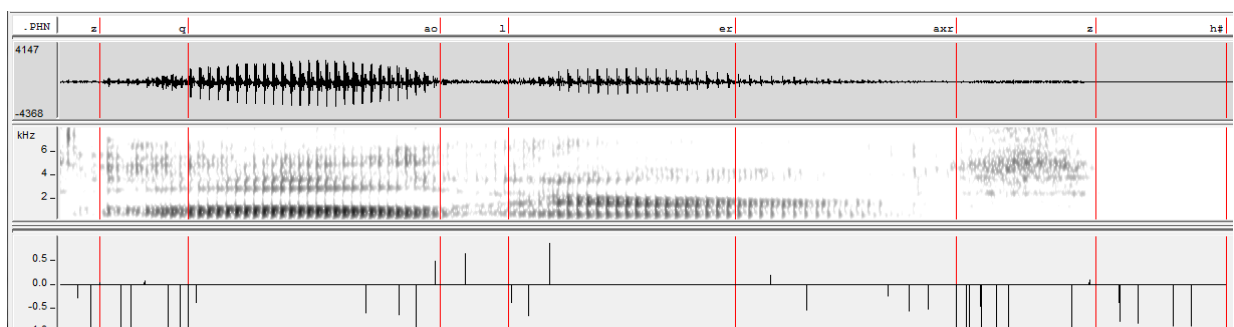


圖 5.5：音素端點偵測前後音素為母音之範例，由上至下分別表示音素層級之人為時間標記的文字轉寫、語音信號、聲譜圖、音素端點偵測器輸出之概似度

### (4) 前後相鄰音素為靜音

靜音通常表示無任何語音信號的區段，但在 TIMIT 語料庫語句內的某一字詞音素與音



素間的短停頓則以 epi 表示。圖 5.6 舉出前後音素為 (tcl-h#) 皆屬於靜音的分類，同前後音素為鼻音的情形相似，僅觀察語音波形也亦難標記正確的音素端點位置，為造成前後音素為靜音時偵測漏失率較高的原因。由音素端點偵測器輸出概似度的觀察中，在 (tcl-h#) 音素的區間中音素候選端點之概似度同樣非常地低，顯示出偵測器偵測不出這些候選端點是音素的端點，藉由調整臨限值也亦難偵測出音素端點。

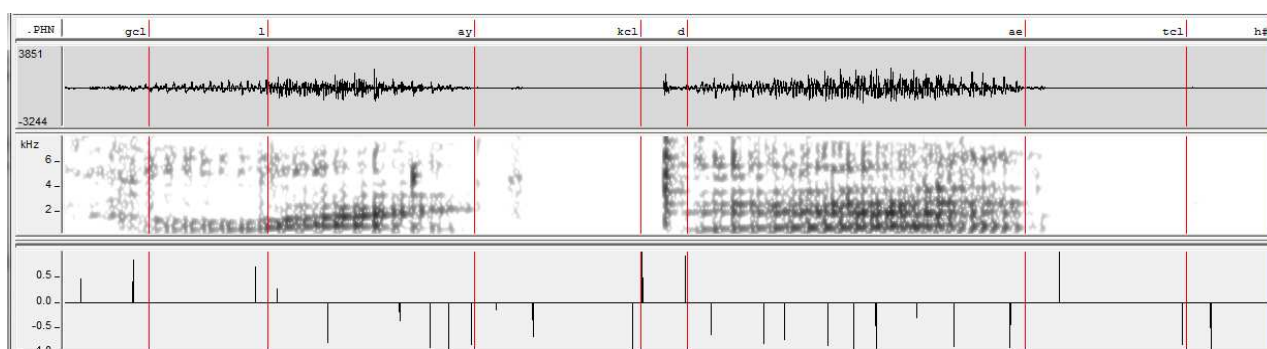


圖 5.6：音素端點偵測前後音素為靜音之範例，由上至下分別表示音素層級之人為時間標記的文字轉寫、語音信號、聲譜圖、音素端點偵測器輸出之概似度

### ► 誤報率分析

由先前所述前後音素為不同發音方法的偵測漏失率較低，但誤報率與偵測漏失率是成反比的，亦即在不同的發音方式的轉換期間語音信號頻譜的劇烈變化容易產生誤報的情形，然而本論文以取樣點式聲學參數挑選音素候選端點的方式與傳統音框解析度對照之下，在此情況卻是更加容易產生較多的音素候選端點，可能造成誤報率增高的情形。故以下分析在前後音素為不同發音方法時誤報率的差異並作討論。表 5.5 為基於已偵測到之邊界其相鄰音素為不同發音方法的統計資料，實驗結果依發音方法之分類顯示。在表中流音以及母音等發音方式之邊界，有較高的誤報率，而其他發音方法誤報率並沒有明顯偏低的情形。此種現象顯現了在發音方法轉換時均有相當地比例產生誤報。藉由實驗結果對應於語音信號以及聲譜圖的觀察，最易發生誤報率的原因有以下幾種：

#### (1) 語音波形不完整

語音波形因語者聲音產生沙啞或是發音不連續造成波形破碎，使得誤報多出現在音素的

尾部。

### (2) 音素對應至頻譜之變化

某些音素因為發音位置或是方法的變化，使得頻譜上的變化使偵測器產生誤報。如圖 5.7 之範例，在 (ow) 的音素之內，發音位置的變化造成該音素頻譜產生改變造成誤報。

### (3) 錄音之雜音

錄音雜音如呼氣聲及發聲之雜音等等，因為接續雜音後便是正常語音，此種情形亦會造成誤報產生。

表 5.5：TIMIT 測試語料中相鄰音素為不同的發音方法之誤報率

Manners	FA rate	Manners	FA rate
Stop	16.51%	Nasals	14.11%
Affricates	14.01%	Glides	17.40%
Fricatives	15.06%	Vowels	16.83%
		Silence	14.40%

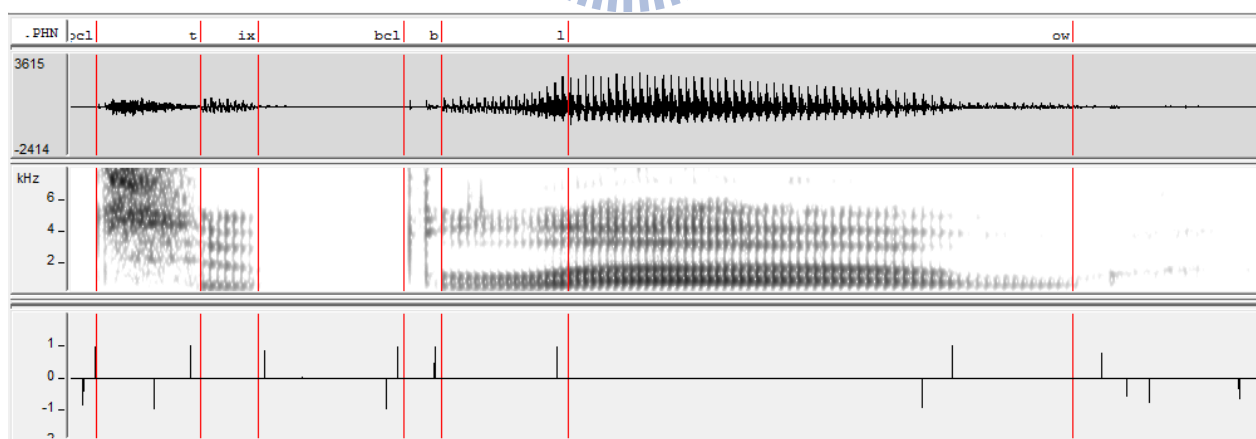


圖 5.7：音素端點偵測誤報率分析之範例，由上至下分別表示音素層級之人為時間標記的文字轉寫、語音信號、聲譜圖、音素端點偵測器輸出之概似度

## 5.2 使用國語語料庫之實驗結果

### 5.2.1 TCC300 語料庫實驗結果分析

實驗結果以音框式抽取參數的 HMM 架構，作強迫切割所獲得的類音素層級初始自動分段位置來比較，並觀察本論文自動分段位置之精準度是否有進一步地提升。由第四章所述，在得到對應類音素層級之音素端點偵測器後，將 HMM 的類音素自動分段結果以端點偵測器所產生的概似度經正規化後作為分數，進行維特比搜尋並限制搜尋範圍在初始位置前後 100 毫秒之內，最後得到本論文之類音素層級自動語音分段結果。

首先，以下列舉兩個語音波形比較音素端點與 HMM 的自動分段位置。由下列圖 5.8、5.9 之中，可由方圈之圈選處之音素端點位置觀察到，無論是音節與音節之間的短停頓或是聲母與韻母之間的端點位置都非常準確，尤其是母音和塞擦音、摩擦音之間的邊界端點與 HMM 之分段位置相比確實精準許多，而在聲譜上觀察這些端點位置可看出頻譜分佈差異極大，亦是正確的端點位置。圖 5.8 所示之方圈圈選處，我們亦可發現在母音轉變至鼻音韻尾的情形，其音素端點位置之準確度仍能保持良好的水準；而在爆破音前的短停頓亦能調整至適當的端點位置。由上述實驗結果在語音波形的觀察下，顯示了取樣點式聲學參數對 HMM 之自動分段位置做修正後，其自動分段之效能確有提升。

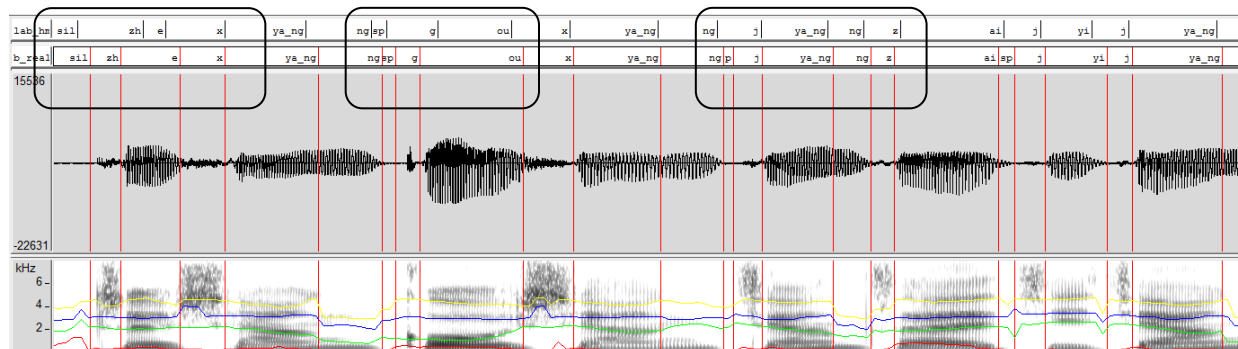


圖 5.8：國語語句自動語音分段之範例一，由上至下的圖形分別表示 HMM 分段位置及音素端點偵測之分段位置、語音波形、聲譜圖

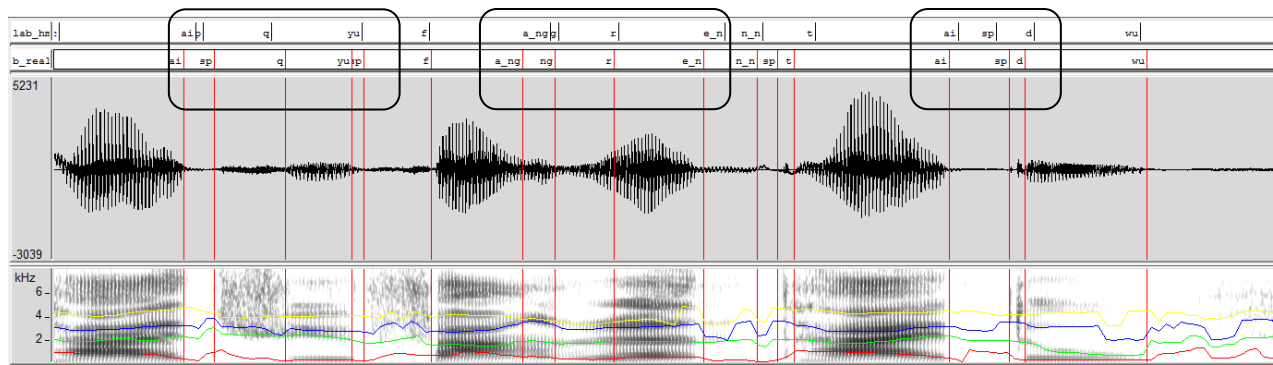


圖 5.9：國語語句自動語音分段之範例二，由上至下的圖形分別表示 HMM 分段位置及音素端點偵測之分段位置、語音波形、聲譜圖

接著，分別在成功大學與交通大學各隨機選取 7 段，共 1698 個音節，作人為標記的標準答案。統計 HMM 自動分段位置和實驗結果對人為標記的端點位置的誤差並以不同絕對偏差值之包含比率來表示，如圖 5.10。圖中以每 5 毫秒為一單位，本論文所提出之方法在 15 毫秒內之邊界包含比率中，可明顯看出與 HMM 自動分段結果的差距，在 5 毫秒內即可達到 46% 的包含比率。此數據顯示本研究方法能有效地改正原本 HMM 的自動分段結果，提升自動語音分段的精確度。在另一方面，隨著與人為標記位置的誤差增大，兩者方法之間的差距慢慢地縮小，在絕對偏差值 30 毫秒的範圍之後仍還有約 10% 的邊界誤差極大，以致於無法涵蓋其中，而本研究方法在其範圍之後效能與 HMM 相比甚至較差，其原因歸類於下列：

### 1. 連續語音所產生的一些現象

首先，連音現象易使得兩者實驗方法皆難以判斷端點位置。例如「第 (d-e) - (yi)」中 (e-yi) 的端點位置，發音方法與口型上的變化都相似而頻譜亦趨於平滑變化，造成端點位置判定上的困難。接著，聲調語言是以音節為發音基礎，但實際上對應於語音信號的音素層級文字轉寫仍會有所差異，如發音位置同化的現象。

### 2. 語料庫錄音的背景雜訊

語料庫中不穩定之錄音品質，造成有部分音檔的背景雜訊過大，在取樣點式聲學參數之子頻帶信號波封反映出劇烈變動的情形，因此造成端點位置標記產生偏差。

### 3. HMM 自動分段結果與人為標記之間誤差過大

由第四章所述，本研究之自動語音分段方法是基於 HMM 之自動分段結果再使用端點偵

測器所提供之分數進行維特比搜尋。因此，起始分段位置之誤差過大亦難在搜尋空間找到最佳的候選端點，使得端點位置產生偏差。

#### 4. 類音素音節結構與候選端點個數在該音段過少所引起端點位置標記誤差偏大的情形

由於本論文是將韻母定義為介音以及韻腳除去鼻音韻尾後所組成，但是在韻母音中雙母音中的音素的變化卻是容易造成本研究方法的端點位置標記誤差增大，例如「作 (zuo) 為 (wei)」，韻母 (wei) 中即可分為介音 wu、主元音 ei 和韻腹 eh 以及韻尾 yi，其中在聲譜圖內介音至主元音的變化卻是較 (o-wu) 變化明顯。然而候選端點在這些變化較為明顯的地方容易挑選出來，進而使標記位置錯誤。

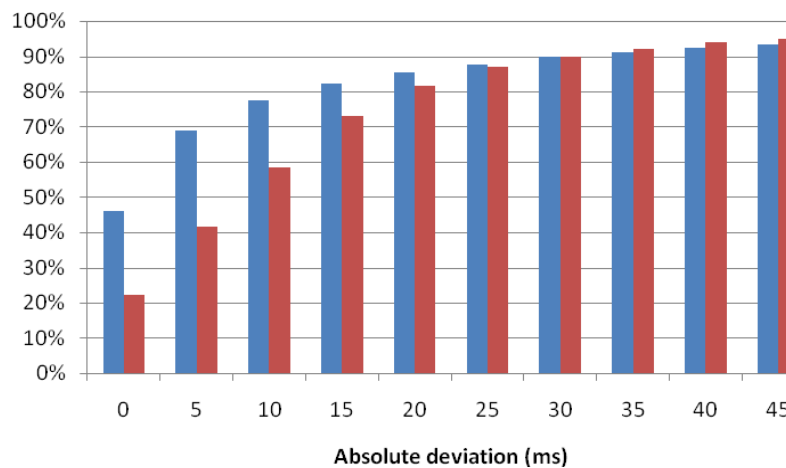


圖 5.10：實驗方法與人為標記位置之誤差在不同絕對偏差值的包含比率直方圖，藍色線(左側)為本論文所提出之方法，紅色線(右側)為使用 HMM 之初始自動分段位置

然而圖 5.10 為實驗結果所有邊界端點與人為標記位置的統計，以下本論文依續 5.1 節音素端點偵測應之誤報率與偵測漏失率的分析結果，將實驗結果依不同發音方法所對應的包含比率做比較，觀察自動語音分段之效能好壞。首先由圖 5.11 中在絕對偏差值為 15 毫秒的範圍內，圖(a)的整體曲線一開始便急遽拉升至近 80% 以上的包含比率，但在圖(b)包含比率之整體曲線則是呈現相較緩慢速度的提高。在圖(a)中，由摩擦音與塞擦音的包含比率相較於圖(b)之結果差距逾 40%，代表著本研究方法確實有助於對此類發音方法之邊界端點來提升自動分段的準確度。然而圖(a)、(b)的結果中發音方法為靜音之曲線趨勢差異為最大，其中隱含著在 HMM 的自動分段結果中，短停頓不易標記出來抑或是不夠準確的情況，此一現象亦顯



現出本研究方法對於音節間短停頓的修正，有大幅度地改進。

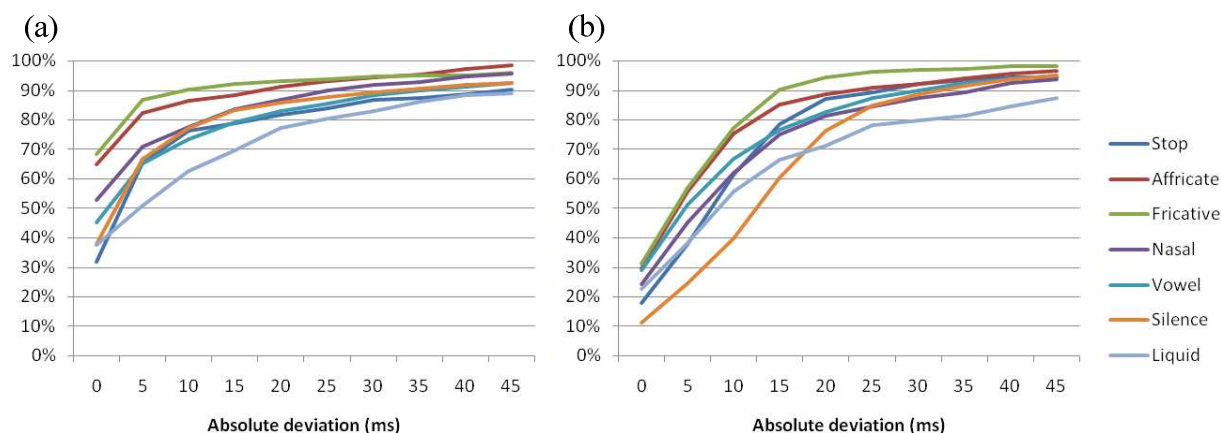


圖 5.11：實驗方法與人為標記位置之誤差以發音方法對應不同絕對偏差值的包含比率直方圖，(a) 本論文所提出之方法，(b) HMM 之初始自動分段位置

## 5.2.2 Treebank 語料庫實驗結果分析

利用第四章所述之自動語音分段的方法，來印證在國語 Treebank 語料庫的效能，並以音節層級和聲/韻母層級訂定目標函數進行自動語音分段，最後得到對國語 Treebank 語料庫兩種不同層級的自動分段實驗結果。

Treebank 語料庫有人為時間標記的資訊，首先針對音節層級之實驗結果統計 HMM 自動分段位置和實驗結果對人為標記的端點位置的誤差在不同絕對偏差值之包含比率，如圖 5.12。在音節層級方面，本論文所提出之方法相較於 HMM 自動分段結果之準確率也有大幅度地提升。此數據顯示本研究方法能有效地改正原本 HMM 音節層級的自動分段結果，提升自動語音分段的精確度效能。

另一方面，在實驗結果的分析上同樣發生與 TCC300 語料庫相同的問題，圖 5.12 以絕對偏差值為 30 毫秒做為分界點，可以觀察出分界點左邊的包含比率相較於分界點右邊的上升幅度較大，這種現象表示出在本研究方法確實能將 HMM 自動分段位置調整至更精確，但分界點右邊代表越難找到一個合適的端點位置使得偏差值的增高，造成分界點右邊包含比率上升幅度趨緩的原因。



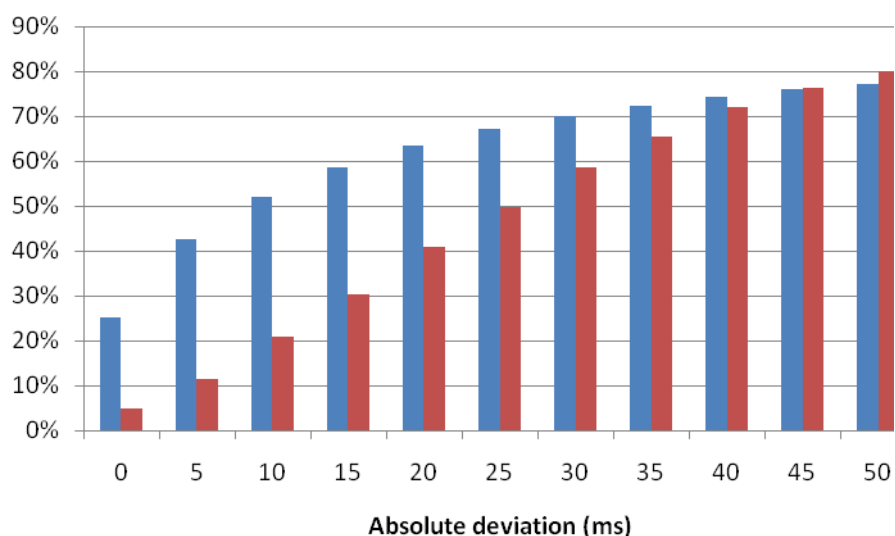


圖 5.12：實驗方法與人為標記位置之誤差在不同絕對偏差值的包含比率直方圖，藍色線(左側)為本論文所提出之方法，紅色線(右側)為使用 HMM 之初始自動分段位置

本論文同樣將聲/韻母層級的端點測器以修正 HMM 自動語音分段之實驗結果顯示了與 TCC300 語料庫類音素層級以及 Treebank 語料庫音節層級實驗相似之實驗結果，在此不多作敘述。而將音節層級與聲/韻母層級之實驗結果相比，可以發現左側音節層級和右側聲/韻母層級有一段包含比率的差距，且聲/韻母層級之實驗結果較佳，如下圖 5.13 所示。此圖凸顯出不同音節結構的層級對自動語音分段效能的影響，其原因將在以下做討論。

由於音檔抽取之取樣點式聲學參數以及挑選音素候選端點的過程為相同步驟，則影響效能差異的關鍵即是對候選端點依照不同層級音節架構所標記之目標函數，而目標函數中所代表不同分類之間的轉移狀態即為自動分段所能調整的端點位置。由第四章所述，對描述語音單元中邊界端點轉移狀態的定義分別由音素端點偵測只有一種描述音素邊界端點的轉移狀態；音節層級轉移狀態有兩種；聲/韻母層級則有四種；最後類音素層級共有五種描述語音邊界端點轉移狀態的目標函數。然而，轉移狀態的個數也象徵候選端點對描述語音單元邊界端點的分類，與只有一種描述音素邊界端點轉移狀態相比，若邊界端點的類型適當地增多便能顯現各轉移狀態統計特性的差異，減低因輸入候選端點之聲學參數特性相似讓端點偵測器產生混淆的可能性。

語音之結構亦隱含前後轉移狀態之間的順序關聯性，以聲/韻母層級舉例，目前候選端點為聲母的狀態，則下一個候選端點就只能為聲母至韻母的轉移狀態或仍是聲母的狀態而不會跳過結構中的分類。在另一方面，此順序之關聯性也可能造成分段位置之絕對偏差值增大，如類音素層級之鼻音韻尾分類，欲在韻母狀態之後尋找最佳韻母至鼻音韻尾轉移狀態之候選端點，但語音信號卻有鼻音弱化的現象，使得維特比搜尋在該音段選擇轉移狀態相對較大的端點造成自動分段效能變差的情形。

因此綜合以上所述且考量語音信號所挑選出候選端點數目以及觀察候選端點之位置，自動語音分段屬聲/韻母與類音素之層級較為合適，也就是圖 5.13 聲/韻母層級之實驗結果較佳的原因。

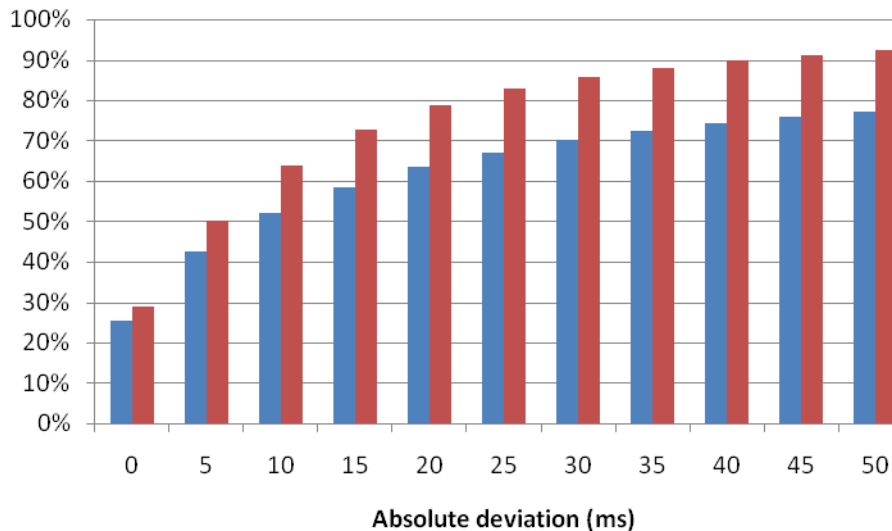


圖 5.13：不同音節結構實驗結果與人為標記位置之誤差在不同絕對偏差值的包含比率直方圖，藍色線(左側)為音節層級，紅色線(右側)為使用聲/韻母層級

在此可作延伸討論，實驗結果顯示對候選端點經過適當地分類標記，可有助於實驗效能之提升。但換句話說，候選端點間之音段也同樣經過了分類標記，那麼以音段為基礎的聲學特性來建立分類的模型，即可應用至語音屬性偵測甚至是語音辨認中。

## 5.3 使用客語四縣語料庫之實驗結果

### 5.3.1 音素端點偵測實驗結果

利用 TIMIT 英文語料庫所訓練的音素端點偵測器來偵測客語語料庫內語句的音素端點。圖 5.14 為偵測客語語句音素端點的範例，在語音波形、聲譜圖與端點偵測器輸出音素候選端點之概似度（範圍在-1~1 之間，其值為 1 則代表為音素端點可能性為最大；相反地，值為-1 可能性最小）的對應觀察中，偵測的端點位置亦可對應至語音波形轉換或是頻譜變化的位置。利用英文語料庫所訓練的模型對客語語料進行音素端點偵測之方式，由觀察實驗結果顯示音素端點偵測確實是可跨語言的。然而，由目前偵測器的輸出結果發現當語音屬性不同的轉換時如發音方法不同，則偵測器在該候選端點產生概似度較高的現象，此種現象也呼應了 5.1 節前後音素為不同發音方法其偵測漏失率相對較低。最後，可藉由適當地調整臨限值來達到偵測器最佳輸出之端點偵測結果，依照此結果我們可將語音信號分為一段段的音段，且這些音段即呈現語音信號中較為穩定的部分，如圖中箭頭所例，亦可提供語音屬性偵測的應用甚至是語音辨識所使用。

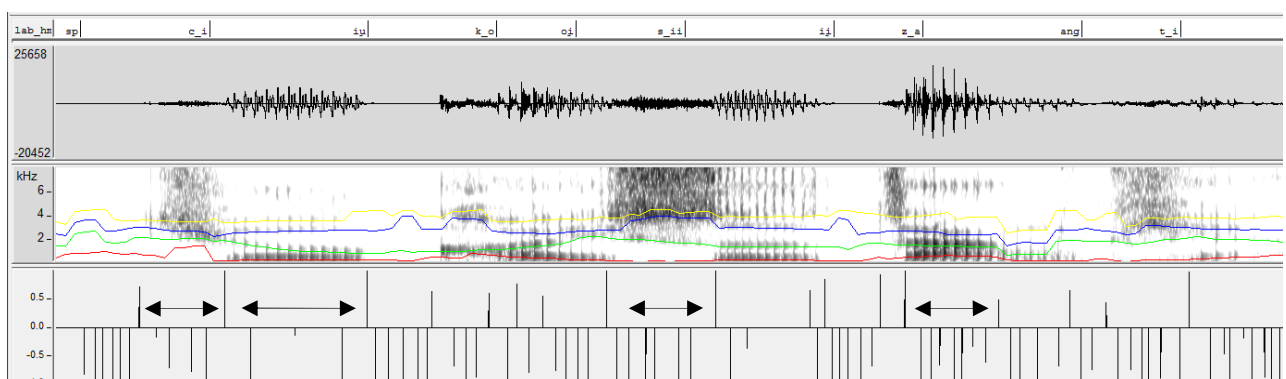


圖 5.14：偵測客語語句音素端點之範例，由上至下的圖形分別代表 HMM 分段位置、語音波形、聲譜圖以及端點偵測器輸出音素候選端點之概似度

### 5.3.2 自動語音分段實驗結果

利用第四章所述之自動語音分段的方法，來印證在客語語料庫的效能，相比於 TCC300 語料庫其差異在於基本語音單位不同，客語為聲/韻母層級而 TCC300 語料庫為類音素層級。因此目標函數為根據聲/韻母層級所訂定並進行自動語音分段的方法，最後得到對客語語料之自動分段位置。

客語語料庫因無人為時間標記的資訊，故以下列舉兩個語音波形的範例來比較本研究方法與 HMM 自動分段的準確度。圖 5.15、圖 5.16 之中，可由方圈之圈選處觀察到一些現象，其音節之間邊界或是聲母與韻母之間的端點位置也都能修正至較為準確的位置，在爆破音與短停頓的交界或是不同發音方式的轉換點尤其明顯，由上述實驗結果在語音波形的觀察下，顯示了本研究所提出之自動語音分段方法對 HMM 之自動分段位置做修正後，其自動分段之效能亦能有所提升。

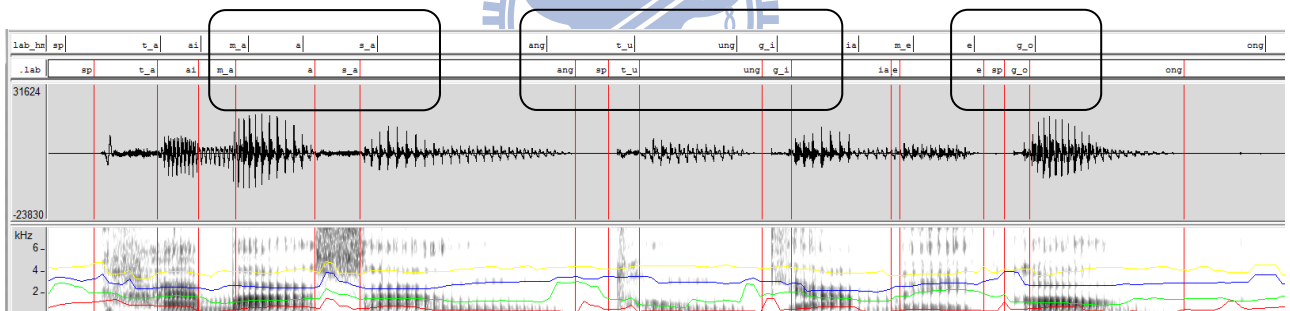


圖 5.15：客語語句自動語音分段之範例一，由上至下的圖形分別表示 HMM 分段位置及音素端點偵測之分段位置、語音波形、聲譜圖

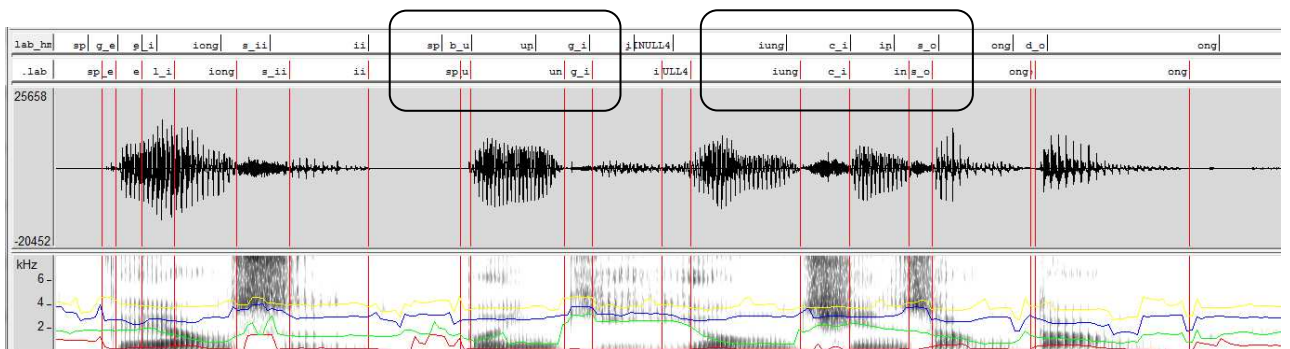


圖 5.16：客語語句自動語音分段之範例二，由上至下的圖形分別表示 HMM 分段位置及音素端點偵測之分段位置、語音波形、聲譜圖

## 5.4 改良頻譜 KL 距離

由先前所述，取樣點式聲學參數是以六個子頻帶信號波封為基礎。然而這些頻帶的頻帶寬度不盡相同，是否會對語音信號的特性產生相關聯的影響；另外一方面，假若依照不同頻帶之貢獻度各加乘上一個具有鑑別性的加權值，能否提升頻譜 KL 距離對評量頻譜差異之效果，故本論文於以下將此問題作延伸探討。

本研究所使用的聲學參數中最能指出語音信號在頻譜間差異的為頻譜 KL 距離，若語音信號在該候選端點之頻譜間的差異越大，其參數值則越高，代表著該候選端點越有可能是不同音素之間的轉換，如(3-6)式。因此，假定我們能增進頻譜 KL 距離的鑑別程度，使得屬於音素端點的候選端點能夠更容易挑選出來並壓抑非音素端點的候選端點，即可更容易且正確地找到屬於音素邊界之候選端點並因此而提升偵測的效能。故本論文參考加權熵[14] (weighted entropy) 的概念，將頻譜 KL 距離進一步改為加權頻譜 KL 距離 (weighted spectral KL distance)，定義為每個子頻帶乘上對應之加權值  $w_i$  的總合，如下式：

$$dw_{KL}[n] = \sum_{i=1}^6 w_i (E_i[n] - E_i[n+1]) \log \left( \frac{E_i[n]}{E_i[n+1]} \right) \quad (5-3)$$

$$= \sum_{i=1}^6 w_i x_{n,i} \quad (5-4)$$

本研究使用最小分類錯誤[15-16] (Minimum Classification Error, MCE) 方法，此方法之目標為期望能夠正確地區別輸入參數向量以達到最佳的辨識或分類之結果，以提升鑑別程度。在此，首先定義觀測向量序列  $X$  為挑選出之所有候選端點序列， $X = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$  且對於第  $n$  個候選端點  $\mathbf{x}_n$  含有六個子頻段信號波封之向量，表示為  $\mathbf{x}_n = [x_{n,1}, x_{n,2}, \dots, x_{n,6}]$ 。而對應不同頻段之加權值  $\mathbf{w} = [w_1, w_2, w_3, w_4, w_5, w_6]$  為實驗欲估測的參數值。考量一組以邊界端點作為分類  $C_m$  的概似度函數  $g_m(X; \mathbf{w})$  如(5-5)、(5-6)式，為觀測向量對應至分類  $C_m$  的概似度。



$$g_1(X; \mathbf{w}) = f(\mathbf{w}^T \mathbf{x}_n) = f(dw_{KL}[n]), \quad X_n \in C_1 \quad (5-5)$$

$$g_2(X; \mathbf{w}) = 1 - f(\mathbf{w}^T \mathbf{x}_n) = 1 - f(dw_{KL}[n]), \quad X_n \in C_2 \quad (5-6)$$

其中分類  $C_1$  表示候選端點為音素端點之類別， $C_2$  則表示候選端點為非音素端點之類別， $f$  為 S 型函數且為對應至值域範圍為[0,1]的連續性函數，並具有可微分之特性，定義如下：

$$f(X) = \frac{1}{1 + \exp(-c(X - x_0))} \quad (5-7)$$

而斜率常數  $c$  可依照參數  $X$  的動態數值範圍大小來設置，其反映出參數數值輸入時 S 型函數所處理的數值範圍，因此觀察原本頻譜 KL 距離的分佈後設定  $c_f$  為 50； $x_0$  表示參數  $X$  欲處理之數值範圍的參數偏權值，初始值為 0.12。

MCE 鑑別式訓練的方法中，可分為三個部分，首先定義鑑別性函數的錯誤分類量測 (Misclassification measure)，接著利用損失函數 (Loss function) 來描述分類判別的錯誤損失，最後是以最小化分類錯誤的目標來估測模型參數。本論文只有使用兩種分類。由上述將錯誤分類量測之數學式定義如下式：

➤ 錯誤分類量測定義為

$$d(X) = -\mathbf{1}(X_n \in C_1)g_1(X_n; \mathbf{w}) - \mathbf{1}(X_m \in C_2)g_2(X_m; \mathbf{w}) \quad (5-8)$$

其中  $\mathbf{1}(\cdot)$  為指標函數，當觀測向量屬於分類  $C_i$  時指標函數值為 1，反之為 0。

➤ 損失函數

損失函數為表示對觀測向量之分類判別錯誤的損失與錯誤量測函數之間的關係，將錯誤量測函數帶入平滑的近似 0-1 損失函數由(5-9)表示：

$$l(X; \mathbf{w}) = l(d(X)) \quad (5-9)$$

其中  $l$  亦為 S 型函數， $x_0$  通常設為 0，而斜率常數  $c_l$  可依照錯誤量測函數  $d(X)$  之動態範圍來



設置，其反映出對分類判別的靈敏度，而由 S 型函數的值域曲線發現，當  $d(X)$  遠小於零，為分類判別正確  $l(X; \mathbf{w})$  沒有判別錯誤造成損失，但  $d(X) > 0$  時則將導致  $l(X; \mathbf{w})$  有明顯的錯誤損失產生。最後以下式(5-10)評量分類器對觀測向量  $X$  的分類判別效能：

$$l(X; \mathbf{w}) = \sum_{n=1}^N l(X_n; \mathbf{w}) \quad (5-10)$$

➤ 最小化分類錯誤的目標

欲得到最佳的加權值  $w_i$ ，本論文以最陡坡降法（Steepest Descent Method）來疊代並求出最佳加權值  $w_i$ 。而在計算之前，先利用變數變換將  $w_i$  作為  $\tilde{w}_i$  的函數如(5-11)式，如此一來便符合所有對加權值  $w_i$  的限制。在此將  $\tilde{w}_i$  的數值初始為 0，使得初始加權值為等機率。

$$w_i = \frac{e^{\tilde{w}_i}}{\sum_{j=1}^6 e^{\tilde{w}_j}} \quad (5-11)$$

接著，將(5-10)式依最陡坡降法進行偏微分的計算，並依照微分連鎖律求取出對應不同頻帶加權值的梯度（Gradient），以批次訓練的方式反覆地疊代更新加權值至收斂，數學式推導過程如以下：

經過一次批次訓練則更新加權值，其公式為：

$$\tilde{w}_i^{(k+1)} = \tilde{w}_i^{(k)} + \mu \cdot \frac{\partial l(X; \mathbf{w})}{\partial \tilde{w}_i^{(k)}} \quad (5-12)$$

其中  $\mu$  為學習速率常數（Step size），並以微分連鎖律對  $l(X; \mathbf{w})$  依照候選端點之類別求取加權值之梯度。

$$\frac{\partial l(X; \mathbf{w})}{\partial \tilde{w}_i} = \frac{\partial l(X; \mathbf{w})}{\partial d} \frac{\partial d}{\partial \tilde{w}_i^{(k)}} \quad (5-13)$$

$$\frac{\partial l(X; \mathbf{w})}{\partial d} = c_i \cdot l(d(X)) (1 - l(d(X))) \quad (5-14)$$

$$\frac{\partial d}{\partial \tilde{w}_i^{(k)}} = -c_f \cdot f(dw_{KL}[n]) (1 - f(dw_{KL}[n])) w_i \cdot (x_{n,i} - dw_{KL}[n]) \quad (5-16)$$

最後之實驗結果我們得到每個頻帶所對應之加權值。如圖 5.17 所示，我們可以看出第一個頻帶的加權值較大，依序為第三個頻帶和第六個頻帶，而第四及第五個頻帶其加權值相對於其他頻帶較小；然而觀察相鄰音素對應發音方法不同的語音信號轉變，若其靜音接至母音、摩擦音接至母音或是塞擦音接至母音，可由看出其頻譜差異最大的是較低頻的部分，這與圖 5.17 的分布一致。重新將加權過後之頻譜 KL 距離用於音素端點偵測器的訓練過程中。

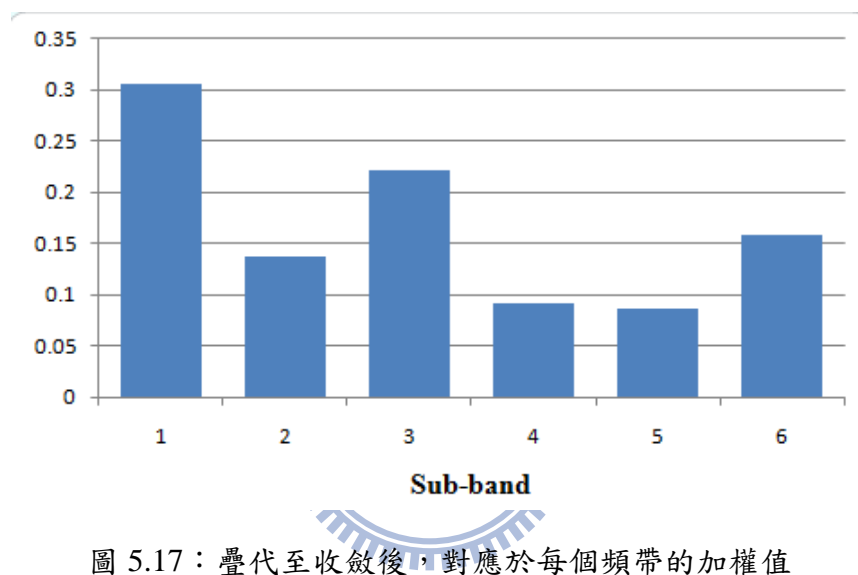


圖 5.17：疊代至收斂後，對應於每個頻帶的加權值

實驗結果由偏權值初始為 0.12 而得，但為了找到適當的加權值以提供重新建立音素偵測器之模型，本論文根據不同之偏權值與調整挑選候選端點之臨限值 (pthD) 之間的定量分析實驗，觀察對每個頻帶加權值的影響。挑選候選端點之臨限值越低則代表挑選出越多的音素候選端點，圖 5.18 顯示了對於訓練加權值的樣本數量 (由 509210 至 677089 個候選端點) 對加權值幾乎無任何影響，而以不同之偏權值主導了各頻帶加權值的分佈情形。

最後整理以不同加權值代入四種挑選候選端點臨限值的實驗結果，當偏權值為 0.12 且臨限值為 0.01 時，與先前之實驗結果相比 EER 可下降約 1.6%，如圖 5.19 所示。

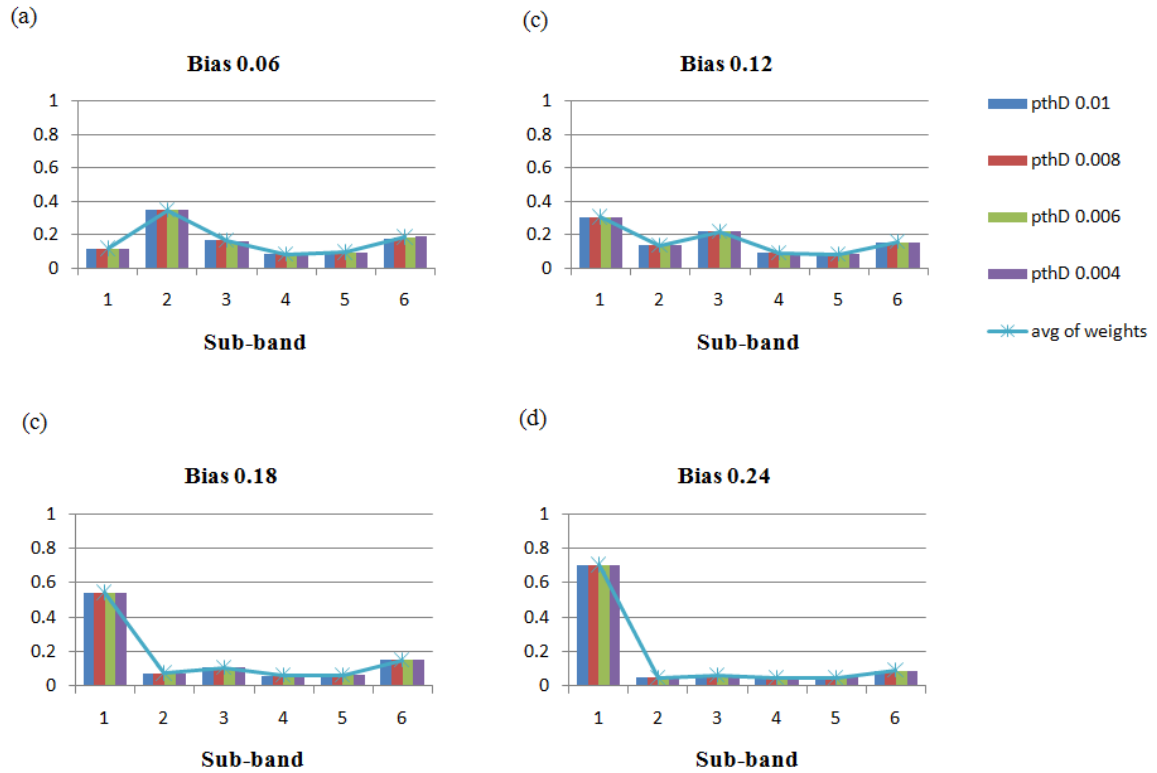


圖 5.18：加權值為根據不同參數偏權值與調整挑選候選端點臨限值的結果，偏權值分別為(a) 0.06 (b) 0.12 (c) 0.18 (d) 0.24

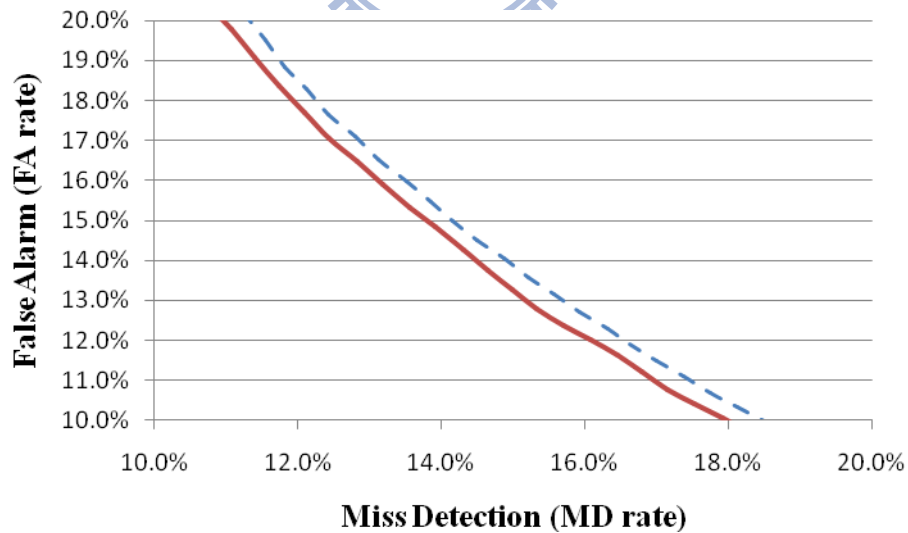


圖 5.19：加入加權頻譜 KL 距離於 TIMIT 測試語料誤報率與偵測漏失率之對應曲線圖，藍色虛線為原本實驗結果曲線，紅色實線為加入加權值實驗結果曲線

# 第六章：結論與未來展望

## 6.1 結論

本論文以獲得一個精確的自動音素端點偵測系統為目標，進行音素端點偵測及自動音素分段的實驗，以提升對應於語音信號之文字轉寫時間標記之精準度，降低人為標記語料庫費時費力的繁雜過程，並期望作為語音辨識或是語音合成系統中處理文字轉寫的標準流程。

本研究提出數個取樣點式聲學參數如各頻段信號波封、聲學參數之上升率、頻譜熵以及頻譜 KL 距離，以描述語音信號中各種不同音素之語音特性，加入音素端點偵測以及自動語音分段的系統架構中，針對音素端點偵測是以音素之邊界端點作為目標函數，另外自動語音分段的架構是根據文字轉寫所使用不同基本語音單位（音節、聲/韻母、類音素）的層級來分類，並依照各個分類彼此之間可能的轉換狀態訂定目標函數，再分別使用類神經網路之多層感知器架構，以半監督式之模型訓練方法建立起音素端點偵測器模型。

在音素端點偵測的效能及錯誤分析的過程中，對於前後音素為相同發音方法時，因為兩者在頻域與時域上的變化不明顯，使偵測器之偵測漏失率會有增高的現象。

在自動語音分段的效能分析中，實驗以隱藏式馬可夫模型以文字轉寫中不同層級的語音單位，來得到初始的分段位置，再加入端點偵測器的模型進行修正而獲得更好的分段位置。此方式和初始分段位置一同和人為之時間標記位置相比後，準確率確實有效地提升，且根據音節不同層級之實驗結果，分析後顯示聲/韻母、類音素層級之效能較音節高。另一方面，也歸納了數個會造成自動語音分段效能降低的現象。

## 6.2 未來展望

本論文提出之取樣點式語音聲學參數用以提供音素端點偵測器模型的建立，此實驗中仍可再進一步改進，譬如說將其他取樣式聲學參數也同樣加入對語音信號表現上具有鑑別性的加權值，或提取更大範圍的聲學參數，例如音段，其可提供較穩定之語音特性來輔助音素端點的偵測；在偵測器架構方面，亦可採用不同的資料分類架構來進行模型的訓練，抑或是置換成對前後語音有記憶特性之遞迴式類神經網路（Recursive Neural Network, RNN）等等方式來建置模型，加強音素端點偵測模型之強健性。在未來，實驗分析將更深入地針對音素之語音屬性進行相似性分析，可經由統計某些特定的音素轉換對的結果，來觀察不同程度之絕對偏差值對音素偵測結果的影響。最後希望藉由本論文所提出一個嶄新音素端點偵測及自動語音分段的經驗，能為新一代語音辨識系統的能力有所提升。



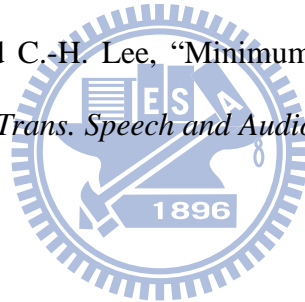
## 參考文獻

- 【1】 F. Malfrère, O. Deroo, and T. Dutoit, “Phonetic alignment: Speech synthesis based vs. hybrid HMM/ANN,” in *Proceedings of the International Conference on Spoken Language Processing*, vol. IV, Sydney, NSW, Australia, 1998, pp. 1571–1574.
- 【2】 Toledano, D.T.; Gomez, L.A.H.; Grande, L.V., “Automatic phonetic segmentation,” *Speech and Audio Processing, IEEE Transactions on*, vol.11, no.6, pp. 617-625, Nov. 2003.
- 【3】 Jen-Wei Kuo and Hsin-min Wang, “Minimum Boundary Error Training for Automatic Phonetic Segmentation,” *The Ninth International Conference on Spoken Language Processing (Interspeech 2006 - ICSLP)*, September 2006.
- 【4】 J.-W. Kuo, H.-Y. Lo, and H.-M. Wang, “Improved HMM/SVM methods for automatic phoneme segmentation,” in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 2057-2060.
- 【5】 K.-S. Lee, “MLP-based phone boundary refining for a TTS database,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 981–989, 2006.
- 【6】 Sorin Dusan and Lawrence Rabiner, “On the Relation between Maximum Spectral Transition Positions and Phone Boundaries,” in *Proc. Interspeech 2006*, pp. 17–21.
- 【7】 Almpanidis, G., Kotti, M., Kotropoulos, and C., “Robust Detection of Phone Boundaries Using Model Selection Criteria with Few Observations,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol.17, no.2, pp.287-298, Feb. 2009.
- 【8】 C.-H. Lee, “From knowledge-ignorant to knowledge-rich modeling : A new speech research paradigm for next generation automatic speech recognition,” *Proc. ICSLP2004*, Keynote speech, 2004.
- 【9】 J. Garofolo et al., “Documentation for the DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM,” Feb. 1993.
- 【10】 Sharlene A. Liu, “Landmark detection for distinctive feature-based speech recognition,”



J. Acoust. Soc. Am. 100 (5), November 1996, pp. 3417-3430.

- 【11】 H. Misra, S. Iqbal, H. Bourlard, and H. Hermansky, “Spectral entropy based feature for robust ASR,” in *Proc. ICASSP 2004*, pp. 193–196.
- 【12】 Jia-lin Shen, Jieh-weih Hung, Lin-shan Lee, “Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments,” in *Proc. ICSLP 1998*.
- 【13】 Nico Tool Kit : Available: <http://nico.nikkostrom.com>
- 【14】 Li Lao, X Wu, L Cheng, X Zhu, “Maximum weighted entropy clustering algorithm,” *Proceedings of the 2006 IEEE International conference on Networking, Sensing and Control*, 1022-1025.
- 【15】 B.-H. Juang, and S. Katagiri, “Discriminative learning for minimum error classification,” *IEEE Trans. Speech and Audio Processing*, vol. 40, no. 12, pp. 3043-3054, Dec 1992.
- 【16】 B.-H. Juang, W. Hou and C.-H. Lee, “Minimum classification error rate Methods for Speech Recognition,” *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 3, pp. 257-265, May 1997.



# 附錄一

TIMIT 語料庫音素分類表

發音方法	標音符號	範例字詞	音素層級之文字轉寫
爆破音	b	bee	BCL B iy
	d	day	DCL D ey
	g	gay	GCL G ey
	p	pea	PCL P iy
	t	tea	TCL T iy
	k	key	KCL K iy
	dx	muddy,	dirty m ah DX iy, dcl d er DX iy
塞擦音	q	bat	bcl b ae Q
	jh	joke	DCL JH ow kcl k
摩擦音	ch	choke	TCL CH ow kcl k
	s	sea	S iy
	sh	she	SH iy
	z	zone	Z ow n
	zh	azure	ae ZH er
	f	fin	F ih n
	th	thin	TH ih n
鼻音	v	van	V ae n
	dh	then	DH e n
	m	mom	M aa M
	n	noon	N uw N
	ng	sing	s ih NG
	em	bottom	b aa tcl t EM
	en	button	b ah q EN
半母音與流音	eng	washington	w aa sh ENG tcl t ax n
	nx	winner	w ih NX axr
	l	lay	L ey
	r	ray	R ey
	w	way	W ey
	y	yacht	Y aa tcl t
	hh	hay	HH ey
hv	ahead	ax HV eh dcl d	

	el	bottle	bcl	b	aa	tcl	t	EL
母音	iy	beet	bcl	b	IY	tcl	t	
	ih	bit	bcl	b	IH	tcl	t	
	eh	bet	bcl	b	EH	tcl	t	
	ey	bait	bcl	b	EY	tcl	t	
	ae	bat	bcl	b	AE	tcl	t	
	aa	bott	bcl	b	AA	tcl	t	
	aw	bout	bcl	b	AW	tcl	t	
	ay	bite	bcl	b	AY	tcl	t	
	ah	but	bcl	b	AH	tcl	t	
	ao	bought	bcl	b	AO	tcl	t	
	oy	boy	bcl	b	OY			
	ow	boat	bcl	b	OW	tcl	t	
	uh	book	bcl	b	UH	kcl	k	
	uw	boot	bcl	b	UW	tcl	t	
	ux	toot	tcl	t	UX	tcl	t	
	er	bird	bcl	b	ER	dcl	d	
	ax	about	AX	bcl	b	aw	tcl	t
	ix	debit	dcl	d	eh	bcl	b	IX tcl t
	axr	butter	bcl	b	ah	dx	AXR	
	ax-h	suspect	s	AX-H	s	pcl	p	eh kcl k tcl t
其他	SYMBOL	DESCRIPTION						
	pau	pause						
	epi	epenthetic	silence					
	h#	begin/end	marker (non-speech events)					
	1	primary	stress marker					
	2	secondary	stress marker					

## 附錄二

### 中文音素分類對照表

表一、國語 21 類聲母表

編號	拼音	注音	編號	拼音	注音	編號	拼音	注音
1	zh	ㄓ	8	g	ㄍ	15	t	ㄊ
2	ch	ㄔ	9	k	ㄎ	16	n	ㄋ
3	sh	ㄕ	10	h	ㄏ	17	l	ㄌ
4	r	ㄖ	11	j	ㄐ	18	b	ㄅ
5	z	ㄗ	12	q	ㄑ	19	p	ㄆ
6	c	ㄘ	13	x	ㄒ	20	m	ㄇ
7	s	ㄙ	14	d	ㄉ	21	f	ㄈ

其中，關於空聲母 (INULL) 為當音節只有韻母發音時給予的預設聲母。

表二、國語 18 類韻母表

編號	拼音	注音	編號	拼音	注音	編號	拼音	注音
1	FNULL1	Φ1	7	a_ng	ㄤ	13	e_ng	ㄥ
2	FNULL2	Φ2	8	o	ㄛ	14	e_n	ㄣ
3	a	ㄚ	9	ou	ㄛㄨ	15	er	ㄝ
4	ai	ㄞ	10	e	ㄝ	16	yi	ㄩ
5	ao	ㄞ	11	eh	ㄝ	17	wu	ㄨ
6	a_n	ㄢ	12	ei	ㄝ	18	yu	ㄩ

其中，關於注音中ㄢ、ㄤ、ㄥ、ㄣ，在本論文中之類音素層級將韻母細分，使鼻音韻尾自成一類。

表三、國語 2 類鼻音韻尾

編號	拼音
1	n_n
2	ng

其中，n\_n 和 ng 分為ㄢ、ㄣ及ㄥ、ㄤ的鼻音韻尾。